The Dissertation Committee for Cesar N. Yahia certifies that this is the approved version of the following dissertation:

Management and Operation of Emerging Mobility Services

Committee:

Stephen D. Boyles, Supervisor

Christian G. Claudel

Gustavo de Veciana

Randy Machemehl

Management and Operation of Emerging Mobility Services

by

Cesar N. Yahia

Dissertation

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
December 2021

Management and Operation of Emerging Mobility Services

by

Cesar N. Yahia

The University of Texas at Austin, 2021

Supervisor: Stephen D. Boyles

Mobility services, such as ridesourcing and e-scooters, are changing the transportation landscape. These services are often unregulated and their impact on traffic or alternative modes is not well understood. In this dissertation, we study two features of mobility services: (1) we analyze the management of ridesourcing platforms through policies aimed at minimizing their congestion externality; in particular, we investigate policies that reduce idling drivers and improve operational efficiency, (2) we explore the interaction between mobility services and alternative modes; in this case, we study the relationship between e-scooter ridership and transit. While the first part of the dissertation develops mathematical models to understand time-dependent policies for ridesourcing management, the second component focuses on statistical and equity analysis of observed e-scooters data in Austin, TX.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mobility services are expanding at a rapid pace. Within the past few years, cities across the globe experimented with a wide range of new mobility options. This growth has been supported by advancement in applications that connect services with users. It is also expected that self-driving technology would reduce operating costs and further accelerate adoption of on-demand mobility services.

That said, will this increase in services translate to the improved mobility of people and goods? In many cities, the current state of unfettered expansion resulted in further congestion, inequity, and pollution. Drivers circle around downtown areas looking for their next passenger, and e-scooters litter public spaces and the environment. Will autonomous taxis increase vehicle miles traveled?, and what proportion of those miles will be without any passengers? Will e-scooters provide accessibility to underserved communities?, or will they be piled up on a downtown sidewalk? Researchers are actively studying all those questions to inform policies that shape well-connected livable cities.

This dissertation investigates strategies for managing ridesourcing services (e.g, Uber/Lyft) by focusing on policies that reduce congestion and minimize operational inefficiencies. Two different ridesourcing management strategies are studied: (1) advanced reservation of rides (supply management) and (2) pricing policies that induce passengers to depart at off-peak periods (demand management). In addition to ridesourcing, the dissertation studies the distribution of e-scooter trips in Austin, TX and how those e-scooters interact with other services such as bus transit.

1.1 Background: Ridesourcing Systems

Recent growth of ridesourcing services is further exacerbating fleet management challenges associated with dynamic and spatially asymmetric passenger demands. Ridesourcing platforms (e.g., Uber and Lyft) need to locate a sufficient number of drivers near anticipated passenger demand to reduce the reach time (i.e., the customer wait time between ride request and the arrival of a driver). However, an abundance of drivers may result in increased driver idle time. This spatiotemporal supply-demand mismatch led platforms to implement a set of strategies aimed at improving operational efficiency.

In general, supply and demand management strategies can be broadly classified into one of the following categories: pricing, fleet sizing, empty vehicle routing (rebalancing), or matching passengers to drivers (Nie, 2017; Zuniga-Garcia et al., 2020). To implement those strategies in practice, the platform uses a set of control levers that include earning guarantees for new drivers, sign-on and added bonuses, and heat maps that show high demand locations where drivers earn more due to surge pricing (Lyft, 2019a,c). In addition, as implemented by Lyft in New York City, platforms can restrict the number of active drivers or force them to drive towards high demand areas if they wish to remain online (Lyft, 2019b).

1.1.1 Modeling frameworks

To theoretically evaluate the impact of demand or supply management strategies, researchers have developed different modeling frameworks that describe ridesourcing systems. Those models often vary depending on the application being studied. However, most models can be classified into one of the following categories: (1) equilibrium-based without stochasticity, (2) steady-state equilibrium analysis of stochasticity, (3) time-dependent without stochasticity, (4) steady-state stochastic analysis within time-dependent

models, (5) time-dependent with transient (non-steady state) analysis of stochasticity. The models proposed in this dissertation fall into the last category, where we develop time-dependent models of ridesourcing systems and evaluate policies using transient analysis of stochastic processes (Yahia et al., 2021b).

Equilibrium analysis methods

The majority of existing studies on ridesourcing systems focus on analyzing interactions between driver supply and passenger demand under *static* equilibrium conditions. These studies seek to evaluate the market share of ridesourcing platforms, competition among platforms, and the impact of ridesourcing platforms on traffic congestion (Bahat and Bekhor, 2016; Ban et al., 2019; Di and Ban, 2019; Qian and Ukkusuri, 2017; Wang et al., 2018; Yahia et al., 2018). Following Yang and Yang (2011), researchers examined the relationship between customer wait time, driver search time, and the corresponding matching rate at market equilibrium (Xu et al., 2020; Zha et al., 2016). Recently, Di et al. (2018) incorporated ridesharing user equilibrium in a network design problem; Zha et al. (2018a) proposed an equilibrium model to investigate the impact of surge pricing on driver work hours; Zhang and Nie (2019) studied passenger pooling under market equilibrium for different platform objectives and regulations; and Rasulkhani and Chow (2019) generalized a static many-to-one assignment game that finds equilibrium through matching passengers to a set of routes. While static equilibrium analysis provides valuable strategic decision-making insights, it fails to address stochasticity and time-dependence in ridesourcing dynamics.

Steady state analysis of stochasticity

To investigate stochasticity in demand/supply management, researchers have developed queueing theoretic models for ridesourcing systems. In particular, closed queue-

ing networks were used to analyze rebalancing and pricing policies (Banerjee et al., 2017; Braverman et al., 2019; Zhang and Pavone, 2016). In these closed queueing networks, the difficulty in designing supply management strategies arises from equilibrium (steady-state) constraints that result in high dimensional non-convex problems (Banerjee et al., 2017). Other queueing based approaches include a double-ended queue to characterize stochasticity in matching (Xu et al., 2020) and an M/G/N queue where each driver is considered to be a server (Li et al., 2019). Spatial stochasticity associated with matching was also investigated using Poisson processes to describe the distribution of drivers near a passenger (Chen et al., 2019; Zhang et al., 2019; Zhang and Nie, 2019).

Those studies focus on steady-state (equilibrium) analysis that disregards the time-dependent variability in demand/supply patterns. Furthermore, temporal variations in demand/supply patterns may occur rapidly, and the system may not attain the steady-state equilibrium conditions (Braverman et al., 2019; Ozkan and Ward, 2020). In addition, policies generated from steady-state optimization in closed queueing networks are open-loop (static); this implies that the policies do not react to the time-dependent stochastic state of the system.

Time-varying models without stochasticity

The importance of time dynamics has been emphasized in recent articles that design time-dependent demand/supply management strategies (Ramezani and Nourinejad, 2018). Wang et al. (2019) proposed a dynamic user equilibrium approach for determining the optimal time-varying driver compensation rate. Similarly, Nourinejad and Ramezani (2020) developed a dynamic model to study pricing strategies; their model allows for pricing strategies that incur losses to the platform over short time periods (driver wage greater than trip fare), and they emphasized that time-invariant static equilibrium models are not capable of analyzing such policies. An alternative dynamic model was

proposed by Daganzo and Ouyang (2019); however, the authors focus on the steady-state performance of their model. While these models can be used to analyze time-dependent policies, the authors do not explicitly consider the spatio-temporal *stochasticity* that results in the mismatch between supply and demand.

Steady-state analysis of stochasticity in time-dependent ridesourcing systems

The most common approach for analyzing time-dependent stochasticity in ridesourcing systems is to apply steady-state probabilistic analysis over fixed time intervals. In other words, a steady state is assumed to be reached within each interval, where parameters such as arrival rate differ across intervals. However, in the context of driver rebalancing, experimental analysis by Braverman et al. (2019) suggests that the time needed to converge to steady-state (equilibrium) in ridesourcing systems is on the order of 10 hours. Thus, since parameters (e.g., passenger arrival rate) vary over much shorter time intervals, the system would not reach the steady-state condition. Another limitation of time-dependent steady-state policies is that they are independent of the realized system state at any time instant. In particular, those policies are based on probabilistic predictions over entire time intervals, and they do not react to the stochastic system state that is realized at a specific time within the time interval.

Transient analysis of stochasticity in time-dependent ridesourcing systems

To address limitations in steady-state methods, Braverman et al. (2019) proposed a time-dependent look-ahead policy that can be used to make rebalancing decisions at any point in time. Recent studies that investigate operational challenges in ridesourcing systems also advocate for transient analysis instead of steady-state models (Nourinejad and Ramezani, 2020; Ozkan and Ward, 2020).

The ridesourcing methods in this dissertation fall into this category of analyzing

time-dependent stochasticity in ridesourcing systems. The proposed policies focus on the *transient* nature of dynamics and do not assume that a steady-state would be achieved. In addition, the proposed state-dependent policies react to the realized fluctuations in the stochastic system state.

1.1.2 Pricing for ridesourcing systems

The majority of existing literature on pricing in ridesourcing systems investigates the role of surge pricing in alleviating or worsening operational inefficiencies. In general, these studies can be classified as either equilibrium-based evaluation of optimal prices or data-driven investigation of pricing inefficiencies.

Modeling ridesourcing systems as two-sided markets, researchers examined the impact of prices on the equilibrium between earning-sensitive drivers and price-sensitive passengers (Bai et al., 2019). In this approach, the prices, demand rate, and expected supply are fixed across different time-periods. Thus, the steady-state equilibrium is assumed to hold within each time period where the optimal price is determined. Alternative steady-state equilibrium methods include: the analysis of threshold-based dynamic pricing strategies, where the prices are determined by the number of idle drivers (Banerjee et al., 2016); spatial pricing across a network of regions (Bimpikis et al., 2019; Zha et al., 2018b); and the use of pricing to alleviate system inefficiencies such as matching drivers to distant passengers at high demand levels (Castillo et al., 2017; Xu et al., 2020; Zha et al., 2018b).

While equilibrium-based methods provide valuable strategic-level insights into supply and demand management (Ban et al., 2019), their value may be limited in operational analysis where the system parameters vary rapidly. As previously mentioned, in the context of driver rebalancing, it was shown that the time needed to converge to a steady-state equilibrium is on the order of 10 hours (Braverman et al., 2019). Thus, since

parameters and system characteristics vary over a much shorter time scale, transient (non-equilibrium) methods are needed for operational decisions. Recently, transient analysis of ridesourcing systems resulted in novel pricing strategies where the platform may incur losses over short time periods (Nourinejad and Ramezani, 2020); the authors emphasize that such policies can not be evaluated using time-invariant steady-state methods.

In addition to model-based analysis, pricing was further examined using data-driven approaches. Notably, by analyzing the spatial variation in the mismatch between supply and demand (search frictions), it was shown that the future earnings of drivers starting at the same location differ significantly based on the assigned destination (Zuniga-Garcia et al., 2020); consequently, there is a need for "destination invariant" pricing mechanisms where drivers starting their trip at the same location and the same time have equal expected future income (Ma et al., 2018). Other data-driven methods include the prediction of future surge pricing patterns to inform driver and rider decisions (Battifarano and Qian, 2019).

As opposed to existing equilibrium-based methods, this research focuses on state-dependent pricing using *transient* analysis of ridesourcing dynamics. In other words, instead of assuming steady-state conditions within successive time periods, we implement real-time pricing that reacts to the current and predicted stochastic system state. Moreover, in contrast to origin-based pricing strategies, the proposed mechanism depends on both spatial and temporal components of the predicted demand.

1.2 Background: E-Scooters

E-scooters offer an alternative travel mode that could reduce congestion and pollution (Gössling, 2020). However, e-scooters have also raised safety concerns as a result of the high injury rate for riders using this mode (Rix et al., 2021). The majority of research

on e-scooters focuses on safety or exploratory travel analysis (Zuniga-Garcia et al., 2021).

In terms of exploratory analyses, Zou et al. (2020) investigated e-scooter data in Washington D. C. It was observed that e-scooters are predominantly used during the evening peak hours and in the middle of the day– indicating that they serve non-commute travel. The authors also noted how e-scooter rides were correlated with events in the area. The trip distances observed were under 1 mile and the median ride duration was around 10 minutes. Thus, similar to observations by Sanders et al. (2020), most e-scooter rides are a replacement for walking. Sanders et al. (2020), who surveyed university staff in Tempe Arizona, noted that barriers to e-scooter ridership include availability of the service and safety concerns.

Safety was especially highlighted in several articles as a key limitation of e-scooters. Yang et al. (2020) mined data from news reports to describe e-scooter incidents. The authors found that the injury rate during the night time was higher and that female riders were less likely to be involved in a fatal crash. As for the injury rate per mile, Rix et al. (2021) found the the e-scooter rate was 175-200 times higher than motor vehicles rates.

This dissertation investigates the relationship between e-scooters and transit. We use CapRemap, Austin's transit network redesign, as a natural experiment to evaluate resulting changes to scooter ridership. The objective of this research is to investigate whether e-scooters can replace transit in areas that lost service.

Chapter 2

Book-Ahead and Supply Management for Ridesourcing Systems

2.1 Introduction

Given the objectives of guaranteeing low customer waiting times and low driver idle time, the following questions arise: how many drivers should a ridesourcing platform supply?, and, how should the platform spatially manage idle drivers based on anticipated demand?

In this chapter, the primary objective is to investigate the role of book-ahead/reserved rides in the management of driver supply. Reservations give precise information characterizing the start time and location of anticipated trips; in turn, the platform can use this information to adjust the availability and spatial distribution of its driver supply. Thus, given a reach time service requirement that the platform seeks to maintain, we analyze the impact of reservations on the number of drivers supplied throughout the network. Moreover, since passengers that schedule a ride in advance expect the driver to arrive within a desired pickup window, our analysis incorporates such priority of book-ahead rides over non-reserved rides.

The proposed supply management framework parallels existing research on ridesourcing systems (Djavadian and Chow, 2017; Lei et al., 2019; Wang and Yang, 2019). The majority of existing studies assume a fixed number of driver supply and/or steady-state (equilibrium) conditions. However, it is increasingly apparent that demand and supply patterns in ridesourcing systems are time-varying. In addition, these variations in demand and supply occur at a fast pace, and the system may never attain a steady state equilibrium.

Thus, our proposed framework for analyzing reservations in ridesourcing systems focuses on the *transient* nature of time-varying stochastic demand/supply patterns. Precisely, for any future point in time, we seek to probabilistically characterize the total number of active (non-idle) drivers; this time-dependent probabilistic characterization is determined by the fraction of book-ahead rides, the stochasticity of non-reserved rides, the anticipated time-varying profile of book-ahead rides, and control policies that aim to maintain reach time priority for book-ahead rides. In more detail, as shown in Figure 2.1, the proposed framework consists of the following three components for managing driver supply:

1. We develop a state-dependent admission control policy that assigns drivers to passengers. The objective of this control policy is to guarantee the reach time service requirement for book-ahead rides. The policy reacts to the realized ride requests and available driver supply. Effectively, the admission control policy ensures that there is a sufficient number of drivers near the location of anticipated book-ahead rides such that the driver can reach the passenger within the pickup window. In other words, the admission control policy ensures that the reach time service requirement is attained for book-ahead rides by choosing which driver to assign to every realized non-reserved ride request.

2. In a predictive approach over an upcoming time-interval, we provide an upper bound on the performance of the state-dependent admission control policy; precisely, the performance of the policy is measured in terms of the probability that the reach time service requirement would be violated for a non-reserved ride. In contrast to steady-state methods, we use *transient* analysis of $M_t/GI/\infty$ to determine the aforementioned upper bound at any point in time throughout the window. In other words, we derive a time-dependent upper bound on the probability of reach

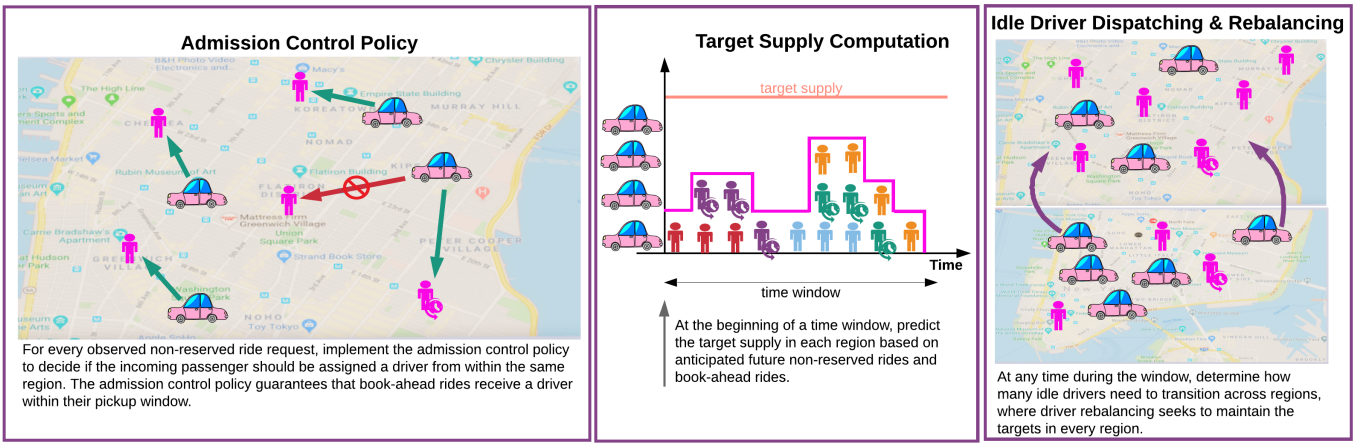

Figure 2.1: Proposed framework for computing the target supply that probabilistically guarantees the reach time service requirement, assigning drivers to passengers to guarantee the arrival of drivers to book-ahead rides within the pickup window, and rebalancing drivers across regions to maintain the targets.

time violation for non-reserved rides. Subsequently, we use the time-averaged value of the upper bound to compute the "target" number of drivers that is required during the upcoming time window; thus, this target limits the probability of reach time service violation to be within a desired performance level.

3. We propose another reactive state-dependent policy for dispatching/rebalancing drivers across multiple regions. In practice, the driver supply may deviate from the predicted target due to the spatiotemporal passenger demand patterns. Thus, we propose a minimum cost flow mechanism that determines the adjustments to the driver supply that are needed to maintain the targets throughout the network. For a specific system state at some time within the time window, the dispatching/rebalancing mechanism determines the number of idle drivers that should transition to adjacent regions to achieve the targets.

The remainder of this chapter proceeds as follows: Section 2.2 describes the proposed model for analyzing time-dependent ridesourcing dynamics. Section 2.3 presents

the admission control policy. Section 2.4 derives an upper bound on the performance of the admission control policy and computes the target supply. Section 2.5 presents the driver dispatching/rebalancing mechanism. Section 2.6 exhibits simulation results using data from Lyft operations in Manhattan. Section 2.7 concludes the chapter.

2.2 System Model

In this section, we describe a general model for time-varying dynamics in rides-ourcing systems. The proposed model represents the number of future *active* rides that initiate in a region. A ride/driver is active from the moment the driver is dispatched to pick up the passenger until the trip is completed. For non-reserved rides, the ride becomes active at the same time as the request is initiated. On the other hand, for book-ahead rides, there is a lag between the time that the request is initiated and the time that the drivers is dispatched to pick up the passenger. While active, drivers are associated with the passenger and can not take on other requests. The ride duration (service time) is the time spent while the driver is active which includes the pick up time. A ride starts when the driver becomes active and ends when the driver is idle again.

The active rides are represented over a set of geographic regions $R = \{1, .., m\}$. These regions are sufficiently small that if a ride request initiates in a region and the assigned driver is operating in the same region, then the reach time is within a desired service level. In other words, if we want the reach time to be under 10 minutes, then the time it takes to drive from any point to any other point within the defined region should be under 10 minutes.

Consequently, we incorporate reservations by providing reach-time priority for book-ahead rides. For a driver to arrive within the book-ahead ride pickup window, the driver must be geographically close to the passenger at the anticipated trip start time.

Table 2.1: Table of notation & definitions

active driver	$\triangleq$	drivers are active from the moment they are dispatched to pick up a passenger and until the passenger leaves the vehicle
idle driver	$\triangleq$	driver waiting to be dispatched (not active)
ride initiation/start	$\triangleq$	time driver is dispatched to pick up passenger
ride completion	$\triangleq$	time passenger leaves vehicle
ride duration	$\triangleq$	total time while driver is active (includes pick up time)
R	$\triangleq$	set of regions $\{1, .., r, .., m\}$
window k	$\triangleq$	time window $(kw, (k+1)w]$
w	$\triangleq$	duration of time window
c_r^k	$\triangleq$	target number of drivers in region r during window k that would probabilistically guarantee a desired reach time service level
$f_r^{P,k}(t)$	$\triangleq$	deterministic process representing active drivers at time $t \in (kw, (k+1)w]$ that are serving requests which initiated in r during *previous* time windows
$f_r^{BA,k}(t)$	$\triangleq$	deterministic process representing active drivers at time $t \in (kw, (k+1)w]$ that are associated with *book-ahead* trips that initiate within window $(kw, (k+1)w]$ in region r
$N_r^k(t)$	$\triangleq$	stochastic process representing active drivers at time $t \in (kw, (k+1)w]$ that are associated with *admitted* stochastic non-reserved rides that initiate within window $(kw, (k+1)w]$ in region r
$\lambda_r^k(t)$	$\triangleq$	demand rate at which stochastic non-reserved ride requests initiate during window k in region r
$g_r^k(\cdot)$	$\triangleq$	probability density function characterizing the ride duration (completion time - trip request time) of stochastic non-reserved rides that appear during window k in region r
$G_r^k(\cdot)$	$\triangleq$	cumulative density function of $g_r^k(\cdot)$
$f_r^{A(\tau_i),k}(t)$	$\triangleq$	active drivers at time $t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]$ corresponding to non-reserved rides that were *previously admitted* between $(kw, \tau_i]$ in region r
τ_i	$\triangleq$	arrival time of the i^{th} non-reserved ride request
D_i	$\triangleq$	ride duration of the i^{th} non-reserved ride
γ_i	$\triangleq$	indicator function/random variable characterizing the event that the i^{th} non-reserved ride request is admitted
B_r^k	$\triangleq$	average blocking probability during window k in region r
δ	$\triangleq$	desired reach time quality of service for non-reserved rides (upper bound on the average blocking probability)
$N_r^{k,\infty}(t')$	$\triangleq$	number of busy servers at time $t' \in (0, w]$ in a transient $\mathrm{M}_t/\mathrm{GI}/\infty$ queue that starts empty at $t' = 0$; equivalently, the number of active non-reserved rides assuming that all stochastic non-reserved requests are admitted
$\rho_r^k(t')$	$\triangleq$	time-dependent mean/variance of the Poisson distribution characterizing $N_r^{k,\infty}(t')$ at time $t' \in (0, w]$
a_r	$\triangleq$	number of active drivers in region r
e_r	$\triangleq$	number of idle drivers in region r
s_r^v	$\triangleq$	virtual supply in region r representing drivers in excess of the target c_r^k that can be removed from region r
d_r^v	$\triangleq$	virtual demand in region r representing drivers that should be added to region r to meet the target c_r^k
Δ_r	$\triangleq$	if region r has virtual demand, then $\Delta_r = -d_r^v$; otherwise, if the region has virtual supply, then $\Delta_r = s_r^v$
h_{ij}	$\triangleq$	recommended driver transitions between region i and j
$\mathbf{1}\{\cdot\}$	$\triangleq$	indicator function or random variable

Thus, we consider that book-ahead ride requests must be assigned a driver from within the same region in which the request initiates, and that satisfying the reach time service requirement for book-ahead rides is equivalent to a driver arriving to the passenger within the pickup window. In Section 2.3, we design an admission control policy that guarantees that book-ahead rides will be assigned a driver from within the same region.

In the proposed ridesourcing model, we do not explicitly analyze ridesharing (i.e., passenger pooling); however, the predicted number of active rides would be a conservative estimate on the corresponding value in ridesharing systems. Furthermore, for tractable target computations, we examine each region separately. In other words, the admission control and corresponding targets assume passengers remain within the zone, disregarding the variation in destinations. Then, to account for the spatial distribution of passenger destinations and the associated movement of drivers across regions, we implement a min-cost flow rebalancing method that maintains the targets across regions. Note that the targets themselves represent a desired number of drivers that is determined by passenger demand; this implies that the targets do not depend on the stochasticity of drivers entering and exiting the system.

We proceed by describing the model for active rides in each region. For each region, this model consists of processes representing book-ahead rides and non-reserved stochastic rides. The processes form the basis of subsequent sections that discuss the admission control policy and the computation of targets.

2.2.1 Time-varying profiles representing rides that will be active in the future

In each region $r \in R$, we represent ridesourcing dynamics over future time windows of length w. At the beginning of each window k, corresponding to time interval

$(kw, (k+1)w]$, the ridesourcing platform can characterize three processes (two deterministic and one stochastic) that will be realized during the upcoming window $(kw, (k+1)w]$. The processes represent *active* drivers at time $t \in (kw, (k+1)w]$ that are serving requests initiated within the region.

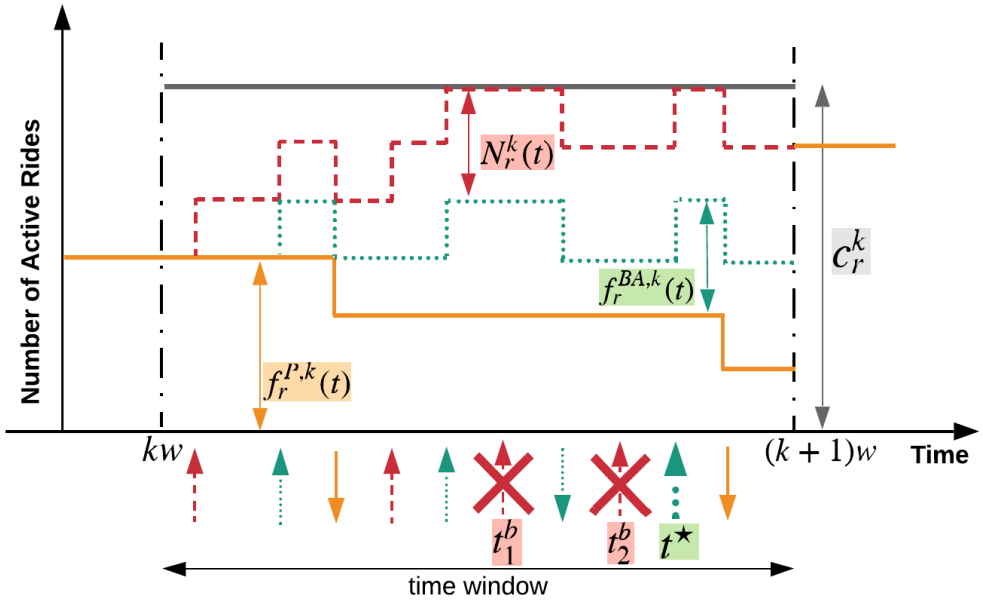

Figure 2.2: System model characterizing the *cumulative* number of rides that will be active in the future at time $t \in (kw, (k+1)w]$. Arrows pointing upwards indicate trip start time. Arrows pointing downwards indicate trip completion. Solid lines correspond to $f_r^{P,k}(t)$, dotted lines correspond to $f_r^{BA,k}(t)$, and dashed lines correspond to $N_r^k(t)$. Non-reserved requests marked with an "X" are blocked requests.

First, we assume that the platform knows the anticipated start time for *book-ahead* rides that will initiate during window k. We also assume that the platform can accurately estimate the corresponding ride duration (i.e. the platform has full trip information for future book-ahead rides). Thus, at the start of window k, the platform can characterize the *deterministic* process $\{f_r^{BA,k}(t) : t \in (kw, (k+1)w]\}$ that represents the number of active drivers at time t associated with book-ahead trips that will initiate in region r within

window k.

Second, at the beginning of time window $(kw, (k+1)w]$, currently active drivers serving rides that started in region r prior to time $t = kw$ are known to the platform. For those *previously observed* trips, we assume that the platform can accurately estimate the trip completion time. Thus, at the start of window k, the platform can characterize the deterministic process $\{f_r^{P,k}(t) : t \in (kw, (k+1)w]\}$. This process represents the number of active drivers at time t that are serving rides that started in region r during previous time windows. In other words, those are previously observed rides that haven't ended yet and may correspond to either passenger type (book-ahead or non-reserved).

Third, at the beginning of window k, the platform also anticipates *non-reserved* stochastic rides that will arise throughout the upcoming window in region r. For those rides, we assume that the platform can estimate the demand (ride request) rate $\{\lambda_r^k(t) : t \in (kw, (k+1)w]\}$. We also assume that the platform can estimate a general distribution $g_r^k(\cdot)$ that corresponds to the ride duration (the CDF of $g_r^k(\cdot)$ is $G_r^k(\cdot)$), and we consider that the duration of any specific non-reserved trip is independent of other trips. Then, we define a stochastic process $\{N_r^k(t) : t \in (kw, (k+1)w]\}$ that represents the number of active drivers at time t associated with *admitted* stochastic rides which initiate in region r during window k. In this case, a non-reserved ride request would be admitted if it is assigned a driver from within the same region.

The deterministic processes $\{f_r^{P,k}(t), f_r^{BA,k}(t) : t \in (kw, (k+1)w]\}$ and the stochastic process $\{N_r^k(t) : t \in (kw, (k+1)w]\}$ are illustrated in Figure 2.2. The figure shows the *cumulative* number of active drivers at time $t \in (kw, (k+1)w]\}$.

The next section describes the admission control policy that decides whether to admit non-reserved rides based on the difference between the predicted targets and the number of active drivers. The admission control policy is state-dependent such that the admission decision is determined for each ride request once the request is observed. In

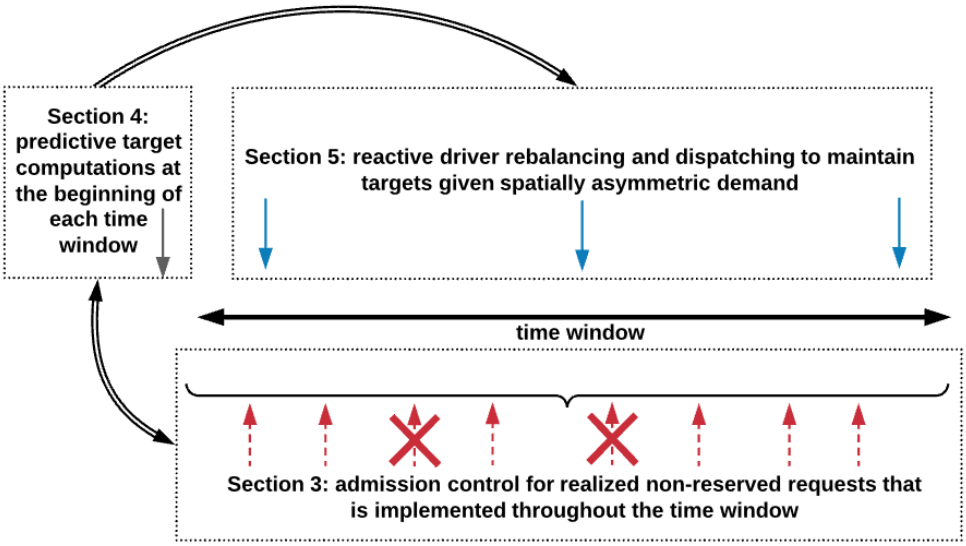

Figure 2.3: Implementation of the proposed framework across time windows.

more detail, the admission decision depends on the current *known* state of the system for the entire duration that the observed ride will be active. Given this policy, we discuss in Section 2.4 how the targets are evaluated at the beginning of the window. However, to compute the targets, we refer to the *predicted* future system state under the control policy, and we resort to a probabilistic characterization of the anticipated non-reserved rides (i.e., we further analyze the stochastic process $\{N_r^k(t) : t \in (kw, (k+1)w]\}$). In other words, the admission control policy uses the targets in determining the *deterministic* admission decisions while the targets are evaluated using the predicted *stochastic* system state that will arise under the control policy. Then, in Section 2.5, we present the driver dispatching and rebalancing mechanism that maintains the targets given the *observed* demand patterns. Figure 2.3 illustrates the relationship between different components of this chapter and the time at which those components would be implemented.

2.3 Admission Control Policy

In this section, we present an admission control policy that is used to assign drivers to realized non-reserved ride requests. In each region, when a non-reserved ride request is observed, the proposed state-dependent control policy determines whether the request should be *admitted* or *blocked*. If the request is admitted, then a driver from within the same region is assigned to serve the request.

The admission decision is based on the supply in the region, the anticipated book-ahead rides, and the previously admitted non-reserved rides. The policy seeks to guarantee that a driver from within the same region would be available to serve anticipated future book-ahead rides. Thus, admission control aims to guarantee that drivers arrive within the pickup window for future book-ahead rides. Since the same policy is implemented for each region, we restrict our discussion in this section to a single region $r \in R$.

At any time $t \in (kw, (k+1)w]$, the admission control policy determines if idle drivers will be available in the region by comparing the number of active rides to the *target supply* c_r^k. The target supply c_r^k, illustrated in Figure 2.2, is the total number of drivers associated with region r during window k; this total includes drivers that are serving ride requests initiated in region r *and* drivers idling in region r. The target c_r^k represents a desired level of driver supply that would probabilistically guarantee the reach time service requirement for non-reserved rides (Section 2.4). The admission control policy assumes that the targets c_r^k will be maintained in each region r throughout the time window k. For tractable computation, the admission control policy also assumes that the passengers destinations remain within the region (in Section 2.5, we devise a driver dispatching/rebalancing mechanism that considers the spatial distribution of demand and seeks to maintain the target across regions).

2.3.1 Policy implementation

A non-reserved ride request is admitted if, upon admission, the total number of active rides does not exceed the target supply for the entire ride duration. Once a non-reserved ride request is observed, the associated ride duration would be also revealed to the platform. Then, there are two cases where the admission control policy would *block* the non-reserved ride request: (1) There are not enough available drivers within the region at the time of request initiation; this is illustrated in Figure 2.2 at time t_1^b, where the sum $N_r^k(t_1^b) + f_r^{BA,k}(t_1^b) + f_r^{P,k}(t_1^b)$ is equal to the target c_r^k. In other words, admission of the non-reserved ride would result in the total number of active rides *exceeding* the target supply at the time of request initiation. (2) Admission of the non-reserved ride would result in reach time service violation for an anticipated book-ahead ride; in Figure 2.2, admission of the non-reserved ride request that initiates at time t_2^b would lead to reach time violation for the book-ahead trip that initiates at $t^\star$ (considering that the observed ride duration of the request that initiates at t_2^b extends beyond $t^\star$). In other words, if the non-reserved ride was admitted at t_2^b, then at $t^\star$ (just before the book-ahead request is anticipated) the sum $N_r^k(t^\star) + f_r^{BA,k}(t^\star) + f_r^{P,k}(t^\star)$ would be equal to the target supply c_r^k; this implies that the total number of active rides would *exceed* the target supply when the book-ahead ride at $t^\star$ starts (equivalently, the book-ahead ride would not be assigned a driver from within the same region).

In more detail, let τ_i be the arrival time of the i^{th} non-reserved ride request, and let D_i be the corresponding ride duration. In addition, let γ_i be an indicator function that takes the value one if the i^{th} non-reserved ride request is admitted. Equation 2.1 gives the expression for γ_i (i.e., Equation 2.1 represents the condition for admission). In Equation 2.1, $f_r^{A(\tau_i),k}(t)$ represents *previously admitted* non-reserved rides that would be active at time $t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]$. In other words, $f_r^{A(\tau_i),k}(t)$ represents previously admitted non-reserved rides that would be active during the time that the i^{th} non-reserved

ride request is being served. Note that the projected ride duration of the i^{th} non-reserved user is restricted to $t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]$ instead of $t \in (\tau_i, \tau_i + D_i]$ since admission control decisions are made per window k (i.e., the rides whose duration extends beyond $t = (k+1)w$ would become part of $f_r^{P,k+1}(t)$).

$$\gamma_i = \mathbf{1}\left\{1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + f_r^{A(\tau_i),k}(t) \leq c_r^k, \quad \forall t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]\right\} \tag{2.1}$$

If we let τ_n and D_n be the arrival time and ride duration of the n^{th} previously observed non-reserved ride (where $n \in \{1, ..., i-1\}$), we can express $f_r^{A(\tau_i),k}(t)$ as shown in Equation 2.2. In this equation, $\mathbf{1}\{\tau_n + D_n > t\}$ takes the value one if the n^{th} previously observed non-reserved ride would be active at time t, and γ_n takes the value one if the n^{th} non-reserved request was admitted.

$$f_r^{A(\tau_i),k}(t) = \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\}\gamma_n, \quad t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}] \tag{2.2}$$

We emphasize that the control policy is state-dependent and applied upon the receipt of each ride request; this implies that the state of the system is deterministic and all the variables (including $\tau_n, D_n, \gamma_n, f_r^{A(\tau_i),k}(t), \tau_i, D_i, \gamma_i$) are known at time τ_i. Then, the admission decision for the i^{th} non-reserved user follows directly from evaluating expressions 2.1 and 2.2.

A non-reserved ride request that is blocked may be assigned a driver from an external region (i.e., the passenger will experience a long wait time). Alternatively, blocked non-reserved requests may be dropped from the system, where this indicates a passenger canceling the ride due to the extended wait time. In the simulation experiments (Section 2.6), we follow the latter approach.

2.4 Target Supply for Probabilistically Guaranteeing the Reach Time Quality of Service

While the admission control policy is a state-dependent policy that is applied during the time window $(kw, (k+1)w]$, it is based on the target supply c_r^k that is determined at the beginning of the time window $t = kw$. For a specific region r, the target c_r^k represents the total number of drivers that is required during window k to probabilistically guarantee the reach time service requirement for non-reserved rides. Drivers are considered to be associated with a region if they are either serving requests that initiated in the region or they are idle within the region. In this section, we discuss how the targets can be computed at the beginning of the time window. First, we derive a time-dependent upper bound on the blocking probability corresponding to the admission control policy. Then, we determine the target number of drivers that limits the time-averaged blocking probability to be below a certain quality of service threshold. In turn, limiting the time-averaged blocking probability is equivalent to limiting the probability of reach time violation for non-reserved ride requests.

In Equations 2.1 and 2.2, representing the admission control policy when the i^{th} non-reserved ride request is received, the values of all the variables are known (for every non-reserved ride request that was previously received, the trip information would have been revealed to the platform). However, at the beginning of the time window, the platform would not know the arrival time, ride duration, and admission decision of a future non-reserved request. Therefore, at the beginning of the time window, $\tau_n, D_n, \gamma_n, f_r^{A(\tau_i),k}(t), \tau_i, D_i, \gamma_i$ are all random variables. To express the probability of admission, we can re-write Equation 2.1 as shown in Equation 2.3. Hence, Equation 2.4

represents the probability that the i^{th} non-reserved ride request would be blocked.

$$P(\gamma_i = 1) = P\left(1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + f_r^{A(\tau_i),k}(t) \leq c_r^k, \quad \forall t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]\right) \tag{2.3}$$

$$\begin{aligned} &P(\gamma_i = 0) = 1 - P(\gamma_i = 1) = \\ &P\left(\exists t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + f_r^{A(\tau_i),k}(t) > c_r^k\right) = \\ &P\left(\exists t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\}\gamma_n > c_r^k\right) \end{aligned} \tag{2.4}$$

Observe that for *predictive* target computations, $f_r^{A(\tau_i),k}(t) = \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\}\gamma_n$ represents stochastic non-reserved ride requests that will be admitted between $(kw, \tau_i]$ *and* will be active at time $t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]$. Recall that future stochastic non-reserved ride requests appear at a demand rate $\{\lambda_r^k(t) : t \in (kw, (k+1)w]\}$ and the corresponding ride durations are generally distributed according to a distribution $g_r^k(\cdot)$. Previously, we defined the stochastic process $\{N_r^k(t) : t \in (kw, (k+1)w]\}$ that represents the number of future active drivers associated with *admitted* non-reserved rides. Notice that $N_r^k(\tau_i) = f_r^{A(\tau_i),k}(\tau_i)$ is the number of admitted non-reserved ride requests that will be active at time τ_i. However, for $t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}]$, $N_r^k(t) \neq f_r^{A(\tau_i),k}(t)$ since $N_r^k(t)$ includes non-reserved ride requests that will be admitted between $(kw, t]$ while $f_r^{A(\tau_i),k}(t)$ is restricted to non-reserved ride requests admitted between $(kw, \tau_i]$.

To determine the target supply c_r^k, we need to evaluate the blocking probability expression in Equation 2.4 for different values of c_r^k. However, this probability expression is difficult to analyze due to the dependence of γ_i (admission of i^{th} non-reserved request) on the random variables τ_n, D_n (arrival time, ride duration) and γ_n (admission) associated with previously arriving non-reserved ride requests $n \in \{1, ..., i-1\}$. In addition, the

arrival time τ_i of the i^{th} non-reserved ride request also depends on the arrival time τ_n of all previous requests. Moreover, the correlations between the random variables have to be considered over the entire time interval $(\tau_i, \min\{\tau_i + D_i, (k+1)w\}]$ and this interval also has time-varying functions $f_r^{P,k}(t)$ and $f_r^{BA,k}(t)$ that impact the admission probability.

Thus, instead of attempting to evaluate Equation 2.4, we provide an upper bound on the blocking probability. In particular, let $\{N_r^{k,\infty}(t) : t \in (kw, (k+1)w]\}$ be the number of busy servers in a *transient* $\mathrm{M}_t/\mathrm{GI}/\infty$ queue that starts empty at the beginning of the window $t = kw$, where the arrivals to the $\mathrm{M}_t/\mathrm{GI}/\infty$ queue appear according to a Poisson process with rate $\{\lambda_r^k(t) : t \in (kw, (k+1)w]\}$ and the service distribution is $g_r^k(\cdot)$.

Theorem 1. *The blocking probability, $P(\gamma_i = 0)$, for the i^{th} stochastic non-reserved ride request that appears at time τ_i is bounded above by* $P\left(N_r^{k,\infty}(\tau_i) \geq c_r^k - \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right]\right)$

Proof. We first start by deriving upper bounds on the blocking probability $P(\gamma_i = 0)$ (Inequalities 2.7–2.9). Then, through Equations 2.11–2.15, we show that the upper bound in Inequality 2.9 can be expressed in terms $N_r^{k,\infty}(\tau_i)$, where $N_r^{k,\infty}(\tau_i)$ is the number of busy servers at time τ_i in a transient $\mathrm{M}_t/\mathrm{GI}/\infty$ queue that starts empty at the beginning of the time window.

$$P(\gamma_i = 0) \tag{2.5}$$

$$= P\left(\exists t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\}\gamma_n > c_r^k\right) \tag{2.6}$$

$$\leq P\left(\exists t \in (\tau_i, \min\{\tau_i + D_i, (k+1)w\}] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\} > c_r^k\right) \tag{2.7}$$

$$\leq P\left(\exists t \in (\tau_i, (k+1)w] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\} > c_r^k\right) \tag{2.8}$$

$$\leq P\left(\exists t \in (\tau_i, (k+1)w] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} > c_r^k\right) \tag{2.9}$$

Inequality 2.7 holds since we are considering that all requests that are received before the i^{th} request are admitted (i.e, $\gamma_n = 1$ for all $n \in \{1, ..., i-1\}$).

Inequality 2.8 holds since we are expanding the time horizon until the end of the window. Inequality 2.9 follows since $\sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} \geq \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > t\}$. Specifically, the number of non-reserved ride requests that are received between $(kw, \tau_i]$ and are still active (being served) at time τ_i is *at least as large as* the corresponding number of non-reserved ride requests that are received between $(kw, \tau_i]$ and are still active at time $t \in (\tau_i, (k+1)w]$ (i.e. $t \geq \tau_i$).

Then, we can rearrange the last expression in Inequality 2.9 as follows:

$$P\left(\exists t \in (\tau_i, (k+1)w] : 1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} > c_r^k\right) \tag{2.10}$$

$$= 1 - P\left(1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} \leq c_r^k, \quad \forall t \in (\tau_i, (k+1)w]\right) \tag{2.11}$$

$$= 1 - P\left(1 + \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right] + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} \leq c_r^k\right) \tag{2.12}$$

$$= P\left(1 + \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right] + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} > c_r^k\right) \tag{2.13}$$

$$= P\left(\sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} > c_r^k - \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right] - 1\right) \tag{2.14}$$

$$= P\left(\sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} \geq c_r^k - \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right]\right) \tag{2.15}$$

Equality 2.12 follows since $f_r^{P,k}(t) + f_r^{BA,k}(t)$ are the only components that depend on t in

expression 2.11, and if the sum $1 + f_r^{P,k}(t) + f_r^{BA,k}(t) + \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\}$ is less than or equal to c_r^k at

$\tilde{t} = \underset{t \in (\tau_i, (k+1)w]}{\arg\max} \left[f_r^{P,k}(t) + f_r^{BA,k}(t) \right]$, then the aforementioned sum is less than or equal to c_r^k for all $t \in (\tau_i, (k+1)w]$.

Equality 2.15 follows since $\sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\}$, $\max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t) \right]$, and c_r^k are all integer values representing the number of active drivers or driver supply.

Thus,

$$P(\gamma_i = 0) \leq P\left(\sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\} \geq c_r^k - \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t) \right] \right) \tag{2.16}$$

let $N_r^{k,\infty}(\tau_i) = \sum_{n=1}^{i-1} \mathbf{1}\{\tau_n + D_n > \tau_i\}$,

Then,

$$P(\gamma_i = 0) \leq P\left(N_r^{k,\infty}(\tau_i) \geq c_r^k - \max_{t \in (\tau_i, (k+1)w]} \left[f_r^{P,k}(t) + f_r^{BA,k}(t) \right] \right) \tag{2.17}$$

$N_r^{k,\infty}(\tau_i)$ represents the number of stochastic non-reserved ride requests that are received between $(kw, \tau_i]$ and are active at time τ_i. Thus, $N_r^{k,\infty}(\tau_i)$ is similar to $N_r^k(\tau_i)$ with the main difference being that $N_r^k(\tau_i)$ is restricted to admitted non-reserved ride requests while $N_r^{k,\infty}(\tau_i)$ accounts for *all received requests* (i.e., $N_r^{k,\infty}(\tau_i)$ assumes that all requests are admitted regardless of the admission control policy). As previously described, stochastic non-reserved ride requests start arriving *after the beginning of the time window* ($t = kw$) according to a Poisson process with demand rate $\{\lambda_r^k(t) : t \in (kw, (k+1)w]\}$ and their ride duration follows the general distribution $g_r^k(\cdot)$. Then, the system corresponding to $N_r^{k,\infty}(\tau_i)$ can be described as a transient $M_t/GI/\infty$ queue that *starts empty* at $t = kw$, receives requests at the rate $\{\lambda_r^k(t) : t \in (kw, (k+1)w]\}$, has a generally distributed service rate $g_r^k(\cdot)$, and has an infinite number of servers (all requests are admitted). In this

context, $N_r^{k,\infty}(\tau_i)$ (the number of active rides at time τ_i) represents the number of busy servers at time τ_i in the transient $M_t/GI/\infty$ queue.

□

Given this upper bound in Theorem 1, we can limit the blocking probability at time τ_i to be below a certain quality of service threshold δ by ensuring that the upper bound is below δ (as shown in Inequality 2.18). Importantly, while $P(\gamma_i = 0)$ is difficult to evaluate as mentioned earlier, the upper bound can be evaluated for any value c_r^k and at any time τ_i using transient analysis of $M_t/GI/\infty$ queues (Section 2.4.1). Subsequently, after illustrating how the upper bound can be evaluated at any time for a specific value of c_r^k, we discuss (Section 2.4.2) how to use this upper bound to determine the target supply, where the target supply is the minimal c_r^k that limits the time-averaged blocking probability to be below the threshold δ.

$$P(\gamma_i = 0) \leq P\left(N_r^{k,\infty}(\tau_i) \geq c_r^k - \max_{t\in(\tau_i,(k+1)w]}\left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right]\right) \leq \delta \tag{2.18}$$

2.4.1 Time-dependent distribution of the number of busy servers in an $M_t/GI/\infty$ queue

To evaluate the upper bound $P\left(N_r^{k,\infty}(\tau_i) \geq c_r^k - \max_{t\in(\tau_i,(k+1)w]}\left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right]\right)$ at time τ_i and for a specific c_r^k, we use a graphical approach that was first recognized by Prékopa (1958) and was subsequently further discussed in articles that analyze $M_t/GI/\infty$ queues (Eick et al., 1993; Foley, 1982). We show that the number of busy servers in an $M_t/GI/\infty$ queue that starts empty, $N_r^{k,\infty}(\tau_i)$, has a *time-dependent* Poisson distribution, and we derive the time-dependent mean associated with this distribution. Thus, since $\max_{t\in(\tau_i,(k+1)w]}\left[f_r^{P,k}(t) + f_r^{BA,k}(t)\right]$ and c_r^k are known values at time τ_i, evaluating the up-

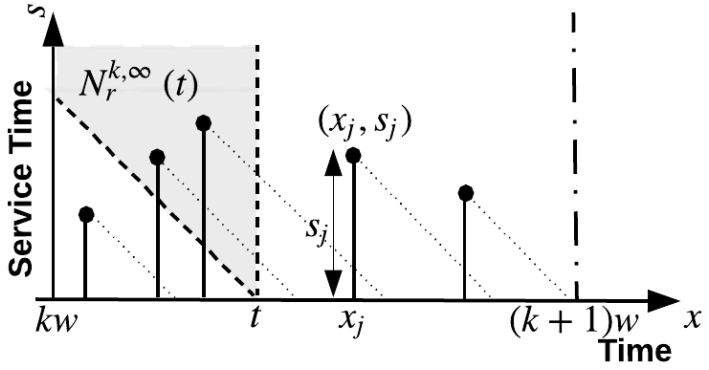

Figure 2.4: Service time vs. arrival time associated with a transient $M_t/GI/\infty$ queue that starts empty at time kw. Since there are an infinite number of servers, all arrivals start being serviced immediately. The dotted diagonal lines represent the decrease in remaining service time as the user is being served. For any time t, the number of users still being served is equal to the number of diagonal lines that intersect a vertical line from t; equivalently, the number of users still being served at t is the number of points in the shaded area.

per bound is equivalent to computing the probability that a Poisson random variable is greater than or equal to a constant.

Referring to Figure 2.4, consider stochastic arrivals to an $M_t/GI/\infty$ queue such that x_j denotes the j^{th} arrival time according to the Poisson process and s_j denotes the corresponding generally distributed service time. In time window $(kw, (k+1)w]$, the $M_t/GI/\infty$ queue is *initially empty* at time kw.

We can think of (x_j, s_j) as a random point in the two-dimensional plane $(kw, (k+1)w] \times [0, \infty)$ that represents the arrival time and service duration. For any two-dimensional set S in $(kw, (k+1)w] \times [0, \infty)$, the number of points in the set represents random sampling of the arrivals Poisson process; thus, the number of points in the set S is *Poisson distributed*. We also know that disjoint two-dimensional sets correspond to independent sampling of a Poisson process; this implies that the number of points in each set is independent of other disjoint sets.

Furthermore, considering an infinitesimal two-dimensional square set with an area

$ds(dx)$, we can see that the mean number of points in that set is $\lambda_r^k(x)dx\left(g_r^k(s)(ds)\right)$; this implies that the intensity of the two-dimensional Poisson distribution is $\lambda_r^k(x)g_r^k(s)$. Thus, the distribution of points defined as (arrival time, service duration) is Poisson over the two-dimensional space, and the *mean* number of points for any set S is given by $\int_S \lambda_r^k(x)g_r^k(s)dsdx$.

To determine the *mean* number of busy servers $\rho_r^k(t)$, we evaluate the integral $\int_S \lambda_r^k(x)g_r^k(s)dsdx$ over the shaded area illustrated in Figure 2.4. This shaded area represents arrivals to the $M_t/GI/\infty$ queue since time kw that have not yet completed at time t. The resulting expression for $\rho_r^k(t)$ is given in Equation 2.19. If we further consider that the arrival rate $\lambda_r^k(x)$ is constant over the time window such that $\lambda_r^k(x) = \lambda_r^k$, the expression for $\rho_r^k(t)$ simplifies as shown in Equation 2.20.

Thus, within each window, $N_r^{k,\infty}(\tau_i)$ is Poisson distributed with a time-dependent mean $\rho_r^k(\tau_i)$. Given a specific value c_r^k, we can use this characterization of $N_r^{k,\infty}(\tau_i)$ to evaluate the upper bound at any time τ_i.

$$\rho_r^k(t) = \int_{kw}^{t}\int_{t-x}^{\infty} \lambda_r^k(x)g_r^k(s)dsdx \tag{2.19}$$

$$\begin{aligned}\rho_r^k(t) &= \int_{kw}^{t}\int_{t-x}^{\infty} \lambda_r^k g_r^k(s)dsdx \\ &= \lambda_r^k\left[t - kw - \int_0^{t-kw} G_r^k(x)dx\right]\end{aligned} \tag{2.20}$$

2.4.2 Target predictions for bounding the time-averaged blocking probability

Knowing that we can evaluate the upper bound on the blocking probability at any time and for any c_r^k, we now investigate the minimal value of c_r^k that limits the *time-averaged* blocking probability to be below a threshold δ. This minimal c_r^k will be referred

to as the *target*, and it represents the number of drivers that the platform seeks to supply during the upcoming time window to limit reach time service violations (i.e., to limit the fraction of non-reserved requests whose reach time will exceed the reach time service requirement).

Precisely, the time-averaged blocking probability in region $r \in R$ during window $(kw, (k+1)w]$ is given in Equation 2.21, where γ_t is an indicator random variable that takes the value one if a passenger that arrives at time t would be admitted. Since Poisson arrivals see time averages (PASTA property), the time-averaged blocking probability is equivalent to the blocking probability of a typical non-reserved ride request that appears between $(kw, (k+1)w]$. Then, the target c_r^k is the desired number of drivers that limits the blocking probability of a typical non-reserved ride request that will appear during the upcoming window.

$$B_r^k = \frac{1}{w} \int_{kw}^{(k+1)w} P(\gamma_t = 0) dt \tag{2.21}$$

As previously mentioned, evaluating the blocking probability in Equation 2.21 is challenging. Thus, to compute the target, we use the time-averaged value of the upper bound in Theorem 1. As shown in Inequality 2.22, if we find the value of c_r^k that limits the time-averaged upper bound to be less than the threshold δ, then this c_r^k will also limit the time-averaged blocking probability to be less than δ. Note that just as we can evaluate the upper bound in Theorem 1 for a specific value of c and at a specific time (Section 2.4.1), we can evaluate the *time-averaged* upper bound for a specific value of c using numerical integration.

$$B_r^k \leq \frac{1}{w} \int_{kw}^{(k+1)w} P\left(N_r^{k,\infty}(t) \geq c_r^k - \max_{\hat{t} \in (t,(k+1)w]} \left[f_r^{P,k}(\hat{t}) + f_r^{BA,k}(\hat{t}) \right] \right) dt \leq \delta \tag{2.22}$$

Therefore, as shown in Equation 2.23, we seek the minimal value c_r^k that restricts B_r^k to be less than or equal to the threshold δ. In Equation 2.23, observe that the time-averaged

upper bound decreases monotonically with increasing values of c; consequently, since c must be a non-negative integer, we can iterate through increasing integer values of c until we find the minimal target c_r^k that ensures that the time-averaged blocking probability is less than δ (alternatively, we may use faster line search techniques).

$$c_r^k = \min_{c \geq 0,\, c \in \mathbb{Z}} \Bigg[c : \frac{1}{w} \int_{kw}^{(k+1)w} P\left(N_r^{k,\infty}(t) \geq c - \max_{\hat{t} \in (t,(k+1)w]} \left[f_r^{P,k}(\hat{t}) + f_r^{BA,k}(\hat{t}) \right] \right) dt \leq \delta \Bigg] \tag{2.23}$$

The targets c_r^k are computed for every region $r \in R$ at the beginning of window k (i.e., at time $t = kw$). If the number of drivers supplied by the platform in each region (either idling in the region or serving requests that initiate in the region) is equal to the corresponding target, then the blocking probability for future non-reserved requests would be less than the threshold δ. Thus, if the targets are provided in each region, the reach time service requirement is probabilistically guaranteed for stochastic non-reserved rides (for book-ahead rides, the reach time service requirement is guaranteed based on the admission control policy in Section 2.3). Apart from target computations, the upper bound on the blocking probability can be used as a performance measure for the admission control policy, where performance of the policy refers to the probability of reach time service violation (for a given level of driver supply).

2.5 Driver Dispatching & Rebalancing Mechanism

In this section, we develop a driver dispatching and rebalancing mechanism that aims to maintain the targets across multiple regions. The targets computed in Section 2.4 represent a desired level of driver supply. In practice, within the time window

$(kw, (k+1)w]$, drivers serving requests that initiated in a region $r \in R$ may finish their trips in other regions. Similarly, drivers serving requests that initiated in an external region $r' \in R \backslash \{r\}$ may finish their trip in region r. Thus, the number of drivers associated with each region may deviate from the corresponding target c_r^k due to observed origin-destination trip patterns. This section presents a dispatching/rebalancing mechanism that computes the minimum number of driver transitions that achieve the targets, where only idle drivers are allowed to transition between adjacent regions. We show that the proposed optimization formulation reduces to a *minimum cost flow* formulation on a transformed network of regions.

In more detail, consider that at some time t the platform aims to determine the necessary driver transitions that maintain the targets. In this section, all the defined variables represent the network conditions at time t; this time t could be either at the beginning of time window $(kw, (k+1)w]$ or within the window. For every region i, let a_i be the number of active drivers serving requests initiated in the region, and let e_i be the number of idle drivers in the region.

In addition, for every region, define a virtual supply s_i^v as shown in Equation 2.24, where the virtual supply represents the number of excess drivers (beyond the target) that can transition to adjacent regions. The virtual supply s_i^v is limited by the number of idle drivers in the region; thus, it is the minimum of the idle drivers e_i and the number of drivers in excess of the target $(a_i + e_i) - c_i^k$. Similarly, define a virtual demand d_i^v as shown in Equation 2.25, where the virtual demand represents the number of additional drivers needed in region i to meet the target c_i^k at time t. Furthermore, for every region i, define Δ_i as shown in Equation 2.26, where Δ_i represents either the demand (expressed as a negative value) or the supply.

$$s_i^v = \begin{cases} \min\left\{e_i, (a_i + e_i) - c_i^k\right\} & \text{if} \quad c_i^k - (a_i + e_i) \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.24}$$

$$d_i^v = \begin{cases} c_i^k - (a_i + e_i) & \text{if} \quad c_i^k - (a_i + e_i) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.25}$$

$$\Delta_i = \begin{cases} -\left[c_i^k - (a_i + e_i)\right] & \text{if} \quad c_i^k - (a_i + e_i) > 0 \\ \min\left\{e_i, (a_i + e_i) - c_i^k\right\} & \text{otherwise} \end{cases} \tag{2.26}$$

For the regions defined in Section 2.2, we construct a directed network $G = (R, E)$. The set of regions R corresponds to the nodes of the network. The set of edges E includes links (i, j) and (j, i) for every pair of *adjacent* regions i and j (see original network in Figure 2.5). Define h_{ij} as the number of drivers that need to transition from region i to the adjacent region j on link (i, j).

The platform rebalancing optimization formulation is shown in Equations 2.27–2.31. In this formulation, the platform seeks to minimize the number of driver transitions (objective 2.27) while ensuring that the targets are maintained (constraint 2.28). In particular, constraint 2.28 specifies that the difference between drivers leaving a region and drivers arriving to a region should match the supply/demand in the region. Constraint 2.29 restricts the number of drivers leaving a region to the number of idle drivers in the region; in other words, this constraint ensures that the optimal solution to formulation 2.27–2.31 (if it exists) describes the number of *idle* drivers transitions to *adjacent* regions (i.e., idle drivers do not transition across multiple regions). The remaining constraints 2.30 and 2.31 ensure that the decision variables h_{ij} are non-negative integers.

$$\min_{h_{ij}:(i,j)\in E} \quad \sum_{(i,j)\in E} h_{ij} \tag{2.27}$$

$$\text{s.t.} \quad \sum_{j:(i,j)\in E} h_{ij} - \sum_{j:(j,i)\in E} h_{ji} = \Delta_i \qquad \forall i \in R \tag{2.28}$$

$$\sum_{j:(i,j)\in E} h_{ij} \leq e_i \qquad \forall i \in R \tag{2.29}$$

$$h_{ij} \geq 0 \qquad \forall (i,j) \in E \tag{2.30}$$

$$h_{ij} \in \mathbb{Z} \qquad \forall (i,j) \in E \tag{2.31}$$

In formulation 2.27–2.31, unless the total supply matches the total demand ($\sum_{i\in R} s_i^v = \sum_{i\in R} d_i^v$) and the network is strongly connected, the optimization problem may not have a feasible solution. Thus, we consider instead the revised formulation 2.32–2.37, where h_i corresponds to drivers added/removed from region i by adjusting the total number of drivers in the network. Since adding or removing drivers would be costly to the platform (e.g., requires incentivizing new drivers or taking drivers offline), we associate a high cost M with such transitions. As a result, in the optimal solution to formulation 2.32–2.37, the total number of drivers is adjusted only if the targets could not be maintained internally via transitions of idle drivers across adjacent regions.

$$\min_{h_{ij}:(i,j)\in E,\, h_i:i\in R} \quad \sum_{(i,j)\in E} h_{ij} + M \sum_{i\in R} |h_i| \tag{2.32}$$

$$\text{s.t.} \quad \sum_{j:(i,j)\in E} h_{ij} - \sum_{j:(j,i)\in E} h_{ji} + h_i = \Delta_i \qquad \forall i \in R \tag{2.33}$$

$$\sum_{j:(i,j)\in E} h_{ij} \leq e_i \qquad \forall i \in R \tag{2.34}$$

$$h_{ij} \geq 0 \qquad \forall (i,j) \in E \tag{2.35}$$

$$h_{ij} \in \mathbb{Z} \qquad \forall (i,j) \in E \tag{2.36}$$

$$h_i \in \mathbb{Z} \qquad \forall i \in R \tag{2.37}$$

In what follows, through a sequence of reformulations, we will show that optimization problem 2.32–2.37 reduces to a minimum cost flow problem on a transformed network.

First, observe that formulation 2.32–2.37 can be rewritten in terms of $h_{i\bullet}$ and $h_{\bullet i}$ that are defined in Equations 2.38 and 2.39. The revised formulation is given in 2.40–2.46. In this case, $h_{\bullet i}$ corresponds to drivers added to region $i \in R$ by adjusting the total number of drivers, and $h_{i\bullet}$ corresponds to drivers removed from region $i \in R$ by adjusting the total number of drivers (i.e., $h_{i\bullet}$ represents drivers that can be removed from the system to avoid having excess idle drivers).

$$h_{i\bullet} = \begin{cases} h_i & \text{if} \quad h_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.38}$$

$$h_{\bullet i} = \begin{cases} |h_i| & \text{if} \quad h_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.39}$$

$$\min_{h_{ij}:(i,j)\in E,\, h_{i\bullet}, h_{\bullet i}: i \in R} \quad \sum_{(i,j)\in E} h_{ij} + M \sum_{i \in R} [h_{i\bullet} + h_{\bullet i}] \tag{2.40}$$

$$\text{s.t.} \quad \sum_{j:(i,j)\in E} h_{ij} - \sum_{j:(j,i)\in E} h_{ji} + h_{i\bullet} - h_{\bullet i} = \Delta_i \qquad \forall i \in R \tag{2.41}$$

$$\sum_{j:(i,j)\in E} h_{ij} \leq e_i \qquad \forall i \in R \tag{2.42}$$

$$h_{ij} \geq 0 \qquad \forall (i,j) \in E \tag{2.43}$$

$$h_{i\bullet}, h_{\bullet i} \geq 0 \qquad \forall i \in R \tag{2.44}$$

$$h_{ij} \in \mathbb{Z} \qquad \forall (i,j) \in E \tag{2.45}$$

$$h_{i\bullet}, h_{\bullet i} \in \mathbb{Z} \qquad \forall i \in R \tag{2.46}$$

Observe that due to the high costs associated with adjusting the total number of drivers, $h_{\bullet i} \leq d_i^v$ for every region i; this inequality implies that the amount of drivers added to region i is less than demand in the region. Similarly, for every region i, $h_{i\bullet} \leq s_i^v$; this inequality implies that the number of drivers disposed from region i (by adjusting the total number of drivers) is less than the virtual supply in the region. If we sum the latter two inequalities over all regions, we get inequalities 2.47 and 2.48. Then, we can rewrite those inequalities using slack variables as shown in Equations 2.49–2.51.

$$\sum_{i \in R} h_{\bullet i} \leq \sum_{i \in R} d_i^v \tag{2.47}$$

$$\sum_{i \in R} h_{i\bullet} \leq \sum_{i \in R} s_i^v \tag{2.48}$$

$$\sum_{i \in R} h_{\bullet i} + \bar{h}_d = \sum_{i \in R} d_i^v \tag{2.49}$$

$$\sum_{i \in R} h_{i\bullet} + \bar{h}_s = \sum_{i \in R} s_i^v \tag{2.50}$$

$$\bar{h}_d, \bar{h}_s \geq 0 \tag{2.51}$$

Intuitively, $\bar{h}_d$ is a slack variable that represents the *demand* that is satisfied through internal driver transitions (as opposed to adding external drivers $h_{\bullet i}$ by adjusting the total number of drivers). Meanwhile, $\bar{h}_s$ is a slack variable that represents the *supply* that is used to satisfy demand through internal driver transitions (as opposed to disposing off

the supply $h_{i\bullet}$ by adjusting the total number of drivers). Therefore, $\bar{h}_d = \bar{h}_s$. A more rigorous approach to show that the equality holds is as follows:

Lemma. $\bar{h}_d = \bar{h}_s = \bar{h}$

Proof. First, we rearrange Equation 2.49 to arrive at Equation 2.52. Then, we can restrict the sum to regions where $\Delta_i < 0$ since by definition $d_i^v = 0$ if $\Delta_i \geq 0$, and since $h_{\bullet i} \leq d_i^v$, then $h_{\bullet i} = 0$ if $d_i^v = 0$ (where $h_{\bullet i} \geq 0$ by definition). Thus, $\Delta_i \geq 0 \Rightarrow d_i^v = 0 \Rightarrow h_{\bullet i} = 0$, and we can restrict the sum to $\Delta_i < 0$ as shown in Equation 2.53.
Equation 2.54 follows by definition of d_i^v and Δ_i when $\Delta_i < 0$.
Equation 2.55 follows by rearranging constraint 2.41. Note that since $\Delta_i < 0$ then $s_i^v = 0$ by definition, and since $h_{i\bullet} \leq s_i^v$ then $h_{i\bullet} = 0$.

$$\bar{h}_d = \sum_{i \in R} d_i^v - h_{\bullet i} \tag{2.52}$$

$$= \sum_{i \in R: \Delta_i < 0} d_i^v - h_{\bullet i} \tag{2.53}$$

$$= \sum_{i \in R: \Delta_i < 0} -\Delta_i - h_{\bullet i} \tag{2.54}$$

$$= \sum_{i \in R: \Delta_i < 0} \left[\sum_{j:(j,i) \in E} h_{ji} - \sum_{j:(i,j) \in E} h_{ij} \right] \tag{2.55}$$

Following a similar approach, we can define $\bar{h}_s$ as illustrated in Equation 2.56.

$$\bar{h}_s = \sum_{i \in R: \Delta_i > 0} \left[\sum_{j:(i,j) \in E} h_{ij} - \sum_{j:(j,i) \in E} h_{ji} \right] \tag{2.56}$$

Then, we can represent the difference between $\bar{h}_d$ and $\bar{h}_s$ as in Equation 2.57.
Observe that if $\Delta_i = 0$, then $\sum_{j:(j,i) \in E} h_{ji} = \sum_{j:(i,j) \in E} h_{ij}$, where this follows by constraint 2.41 ($h_{i\bullet} = h_{\bullet i} = 0$ since $h_{\bullet i} \leq d_i^v$, $h_{i\bullet} \leq s_i^v$ and $d_i^v = s_i^v = \Delta_i = 0$).
Thus, we can rearrange Equation 2.57 to get Equation 2.58.

Then, we can rearrange Equation 2.58 further to get Equations 2.59. Finally, note that $\sum_{i \in R} \sum_{j:(j,i) \in E} h_{ji}$ is a summation over all links in the network, and similarly $\sum_{i \in R} \sum_{j:(i,j) \in E} h_{ij}$ is a summation over all links in the network. This gives Equation 2.60, which proves the lemma.

$$\bar{h}_d - \bar{h}_s = \sum_{i \in R: \Delta_i < 0} \left[\sum_{j:(j,i) \in E} h_{ji} - \sum_{j:(i,j) \in E} h_{ij} \right] - \sum_{i \in R: \Delta_i > 0} \left[\sum_{j:(i,j) \in E} h_{ij} - \sum_{j:(j,i) \in E} h_{ji} \right] \tag{2.57}$$

$$= \sum_{i \in R} \left[\sum_{j:(j,i) \in E} h_{ji} - \sum_{j:(i,j) \in E} h_{ij} \right] \tag{2.58}$$

$$= \sum_{i \in R} \sum_{j:(j,i) \in E} h_{ji} - \sum_{i \in R} \sum_{j:(i,j) \in E} h_{ij} \tag{2.59}$$

$$= \sum_{(i,j) \in E} h_{ij} - \sum_{(i,j) \in E} h_{ij} = 0 \tag{2.60}$$

□

Subsequently, we can add Equations 2.49–2.51 as constraints in formulation 2.40–2.46, where we use $\bar{h} = \bar{h}_d = \bar{h}_s$. The resulting formulation is shown in 2.61–2.71 (Equation 2.50 is first multiplied by a negative sign and then added as a constraint). Note that $\bar{h}$ must be integer since, for each region i, $s_i^v, d_i^v, h_{\bullet i}, h_{i \bullet}$ are all integer.

$$\min_{h_{ij}:(i,j) \in E,\, h_{i\bullet}, h_{\bullet i}: i \in R,\, \bar{h}} \quad \sum_{(i,j) \in E} h_{ij} + M \sum_{i \in R} [h_{i\bullet} + h_{\bullet i}] \tag{2.61}$$

$$\text{s.t.} \quad \sum_{j:(i,j) \in E} h_{ij} - \sum_{j:(j,i) \in E} h_{ji} + h_{i\bullet} - h_{\bullet i} = \Delta_i \qquad \forall i \in R \tag{2.62}$$

$$\sum_{j:(i,j) \in E} h_{ij} \leq e_i \qquad \forall i \in R \tag{2.63}$$

$$\sum_{i \in R} h_{\bullet i} + \bar{h} = \sum_{i \in R} d_i^v \tag{2.64}$$

$$-\left[\sum_{i\in R} h_{i\bullet} + \bar{h}\right] = -\sum_{i\in R} s_i^v \tag{2.65}$$

$$h_{ij} \geq 0 \qquad \forall (i,j) \in E \tag{2.66}$$

$$h_{i\bullet}, h_{\bullet i} \geq 0 \qquad \forall i \in R \tag{2.67}$$

$$\bar{h} \geq 0 \tag{2.68}$$

$$h_{ij} \in \mathbb{Z} \qquad \forall (i,j) \in E \tag{2.69}$$

$$h_{i\bullet}, h_{\bullet i} \in \mathbb{Z} \qquad \forall i \in R \tag{2.70}$$

$$\bar{h} \in \mathbb{Z} \tag{2.71}$$

To map the problem to an equivalent min-cost flow formulation, for each region $i \in R$, we define variables $h_{ii^\star}$ that represent the total number of drivers leaving region i to adjacent regions (Equation 2.72). In addition, for each link $(i,j) \in E$, we define variables $h_{i^\star j} = h_{ij}$. Thus, we can define $h_{ii^\star}$ in terms of $h_{i^\star j}$ as in Equation 2.73. Since h_{ij} is a non-negative integer for all $(i,j) \in E$, we have that $h_{ii^\star}$ and $h_{i^\star j}$ are non-negative integers as well.

$$h_{ii^\star} = \sum_{j:(i,j)\in E} h_{ij} \qquad \forall i \in R \tag{2.72}$$

$$= \sum_{j:(i,j)\in E} h_{i^\star j} \qquad \forall i \in R \tag{2.73}$$

Then, we can express constraint 2.63 in terms of $h_{ii^\star}$ as $h_{ii^\star} \leq e_i$ for all regions $i \in R$. Moreover, we can express the sum of driver transitions across links $(i,j) \in E$ as shown in Equation 2.74.

$$\sum_{(i,j)\in E} h_{ij} = \sum_{i\in R} \sum_{j:(i,j)\in E} h_{ij} = \sum_{i\in R} h_{ii^\star} \tag{2.74}$$

Therefore, we can reformulate optimization problem 2.61–2.71 in terms of the newly defined variables as follows: Substitute Equation 2.74 in the objective function 2.61, replace the sum of drivers leaving a region to adjacent regions with $h_{ii^\star}$ (as in Equation 2.72), replace h_{ij} by $h_{i^\star j}$ and h_{ji} by $h_{j^\star i}$, replace constraint 2.63 with $h_{ii^\star} \leq e_i$, add Equation 2.73 to the constraints, add constraints that restrict $h_{i^\star j}$ to be non-negative integers for all $(i,j) \in E$, and add constraints that restrict $h_{ii^\star}$ to be non-negative integers for all $i \in R$. The revised formulation is shown in 2.75–2.87.

$$\min_{h_{i^\star j}:(i,j)\in E,\, h_{i\bullet}, h_{\bullet i}, h_{ii^\star}:i\in R,\, \bar{h}} \quad \sum_{i\in R} h_{ii^\star} + M \sum_{i\in R} [h_{i\bullet} + h_{\bullet i}] \tag{2.75}$$

$$\text{s.t.} \quad h_{ii^\star} - \sum_{j:(j,i)\in E} h_{j^\star i} + h_{i\bullet} - h_{\bullet i} = \Delta_i \qquad \forall i \in R \tag{2.76}$$

$$\sum_{i\in R} h_{\bullet i} + \bar{h} = \sum_{i\in R} d_i^v \tag{2.77}$$

$$-\left[\sum_{i\in R} h_{i\bullet} + \bar{h}\right] = -\sum_{i\in R} s_i^v \tag{2.78}$$

$$\sum_{j:(i,j)\in E} h_{i^\star j} - h_{ii^\star} = 0 \qquad \forall i \in R \tag{2.79}$$

$$0 \leq h_{ii^\star} \leq e_i \qquad \forall i \in R \tag{2.80}$$

$$h_{i^\star j} \geq 0 \qquad \forall (i,j) \in E \tag{2.81}$$

$$h_{i\bullet}, h_{\bullet i} \geq 0 \qquad \forall i \in R \tag{2.82}$$

$$\bar{h} \geq 0 \tag{2.83}$$

$$h_{ii^\star} \in \mathbb{Z} \qquad \forall i \in R \tag{2.84}$$

$$h_{i^\star j} \in \mathbb{Z} \qquad \forall (i,j) \in E \tag{2.85}$$

$$h_{i\bullet}, h_{\bullet i} \in \mathbb{Z} \qquad \forall i \in R \tag{2.86}$$

$$\bar{h} \in \mathbb{Z} \tag{2.87}$$

Consider the standard minimum cost flow problem given in formulation 2.88–2.90

for a network $G' = (V, A)$ (Ahuja et al., 1993; Wolsey, 1998), where c_{pq} is the cost of a unit flow on link $(p, q) \in A$, x_{pq} are decision variables corresponding to flows on each link $(p, q) \in A$, b_p is the equivalent of supply/demand at node p, and u_{pq} is an upper bound on the flows x_{pq} (i.e., capacity of link $(p, q) \in A$). A necessary condition for feasibility of the optimization problem is $\sum_{p \in V} b_p = 0$.

$$\min_{x_{pq}:(p,q)\in A} \quad \sum_{(p,q)\in A} c_{pq} x_{pq} \tag{2.88}$$

$$\text{s.t.} \quad \sum_{\{q:(p,q)\in A\}} x_{pq} - \sum_{\{q:(q,p)\in A\}} x_{qp} = b_p \qquad \forall p \in V \tag{2.89}$$

$$0 \leq x_{pq} \leq u_{pq} \qquad \forall (p, q) \in A \tag{2.90}$$

Apart from the integrality constraints, the formulation 2.75–2.87 has the same structure as the minimum cost flow optimization problem 2.88–2.90; this implies that the constraint matrix associated with formulation 2.75–2.87 is totally unimodular. Thus, since Δ_i, d_i^v, s_i^v, and e_i are all integer values, each extreme point in the constraint set will be integral. Then, solving the linear programming relaxation in 2.91–2.99 will give us the *integer optimal solution* of optimization problem 2.75–2.87.

$$\min_{h_{i^\star j}:(i,j)\in E,\, h_{i\bullet}, h_{\bullet i}, h_{ii^\star}:i\in R,\, \bar{h}} \quad \sum_{i\in R} h_{ii^\star} + M \sum_{i\in R} [h_{i\bullet} + h_{\bullet i}] \tag{2.91}$$

$$\text{s.t.} \quad h_{ii^\star} - \sum_{j:(j,i)\in E} h_{j^\star i} + h_{i\bullet} - h_{\bullet i} = \Delta_i \qquad \forall i \in R \tag{2.92}$$

$$\sum_{i\in R} h_{\bullet i} + \bar{h} = \sum_{i\in R} d_i^v \tag{2.93}$$

$$-\left[\sum_{i\in R} h_{i\bullet} + \bar{h}\right] = -\sum_{i\in R} s_i^v \tag{2.94}$$

$$\sum_{j:(i,j)\in E} h_{i^\star j} - h_{ii^\star} = 0 \qquad \forall i \in R \tag{2.95}$$

$$0 \leq h_{ii^\star} \leq e_i \qquad \forall i \in R \tag{2.96}$$

$$h_{i^\star j} \geq 0 \qquad \forall (i,j) \in E \tag{2.97}$$

$$h_{i\bullet}, h_{\bullet i} \geq 0 \qquad \forall i \in R \tag{2.98}$$

$$\bar{h} \geq 0 \tag{2.99}$$

The linear program 2.91–2.99 can be mapped to a minimum cost flow program 2.88–2.90 applied on a transformed network illustrated in Figure 2.5. In particular, consider a source node SO where links (SO, i) that connect SO to region $i \in R$ dispatch flows $h_{\bullet i}$. In addition, consider a sink node SI where links (i, SI) that connect region $i \in R$ to SI dispatch flows $h_{i\bullet}$. Let $\bar{h}$ represent the flow between SO and SI. Then, observe that constraint 2.92 is equivalent to constraint 2.89 at all un-starred nodes in the network transformation of Figure 2.5. Similarly, constraint 2.95 is equivalent to constraint 2.89 at all starred nodes. Constraint 2.93 corresponds to constraint 2.89 applied at the source node SO, and constraint 2.94 corresponds to constraint 2.89 applied at the sink node SI. In the network transformation, each link is associated with a $(\text{cost}, \text{capacity})$ label. Observe that the objective function 2.91 can be obtained by plugging the link costs and flow variables in the minimum cost flow objective function 2.88. Also, observe that constraints 2.96–2.99 are the link capacity constraints 2.90 in the transformed network. Furthermore, by definition, $\sum_{i \in R} \Delta_i + \sum_{i \in R} d_i^v - \sum_{i \in R} s_i^v = 0$; this implies that the necessary condition for feasibility in the minimum cost flow program ($\sum_{p \in V} b_p = 0$) is satisfied. Thus, solving the linear program 2.91–2.99 is equivalent to solving the minimum cost flow program 2.88–2.90 using the transformed network.

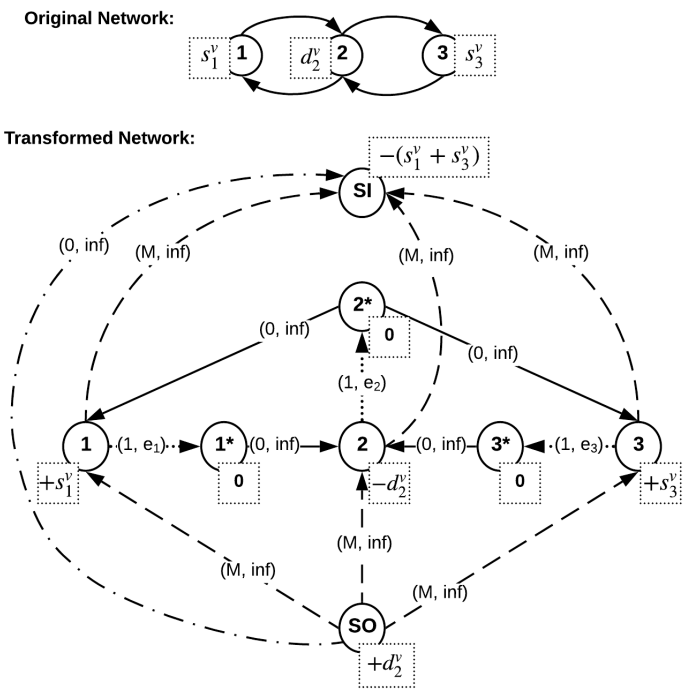

Figure 2.5: Network transformation corresponding to the minimum cost flow program, where solving the integer program 2.32–2.37 using the original network is equivalent to solving the minimum cost flow program 2.88–2.90 using the transformed network. Each link in the transformed network is associated with a $(\text{cost}, \text{capacity})$ label. Each node in the transformed network is either a supply, demand, or transmission node such that values of b_p in constraint 2.89 are within the squares.

Consequently, since the integer program 2.32–2.37 reduces to formulation 2.91–2.99, then solving the integer program 2.32–2.37 on the original network (Figure 2.5) is equivalent to solving the minimum cost flow program 2.88–2.90 on the illustrated transformed network. As a minimum cost flow program, the driver dispatching and rebalancing optimization problem can be solved in polynomial time. The optimal solution of the optimization program represents recommended idle driver transitions that are needed to maintain the targets across regions. Specifically, the optimal solution includes idle drivers

that should transition to adjacent regions *and* idle drivers that should be added to the network by adjusting the total number of drivers in the system. In addition, the optimal solution also includes excess idle drivers that can be removed from the system.

2.6 Simulation Results

In this section, we present experimental results using data from Lyft operations in Manhattan, NYC on Friday December 14th, 2018 (NYCTLC, 2019). We consider trips that started between 16:00–19:00 (local time) in four regions. The regions chosen roughly correspond to four sections of the city as illustrated in Figure 2.6 (1-lower Manhattan, 2-midtown Manhattan, 3-upper west side, and 4-upper east side). For time windows of duration $w = 20$ minutes, we use trip initiation and completion time data available on the New York City Taxi and Limousine Commission website to characterize the processes $\{f_r^{P,k}(t), f_r^{BA,k}(t), N_r^k(t) : t \in (kw, (k+1)w]\}$.

Our primary findings suggest that an increase in the fraction of book-ahead rides leads to a reduction in the total number of drivers that are needed to probabilistically guarantee the reach time service requirement. This reduction in the total number of drivers is also associated with a lower number of idling drivers (i.e., an increase in the driver utilization rate).

2.6.1 System model specification and comparison to observed data

The process $\{f_r^{P,k}(t) : t \in (kw, (k+1)w]\}$ is generated at the beginning of every window k. Specifically, using the available data, $f_r^{P,k}(t)$ represents *previously observed* rides that initiated in region r prior to $t = kw$ and will be active at time $t \in (kw, (k+1)w]$.

To generate the process $\{f_r^{BA,k}(t) : t \in (kw, (k+1)w]\}$ from the New York City data, we randomly sample a fraction p_{BA} of the trips that start during window k in re-

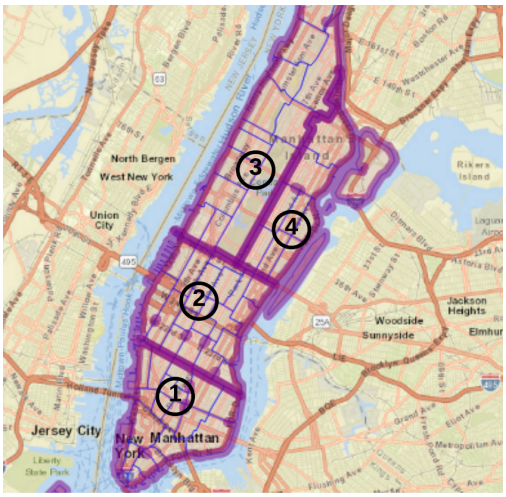

Figure 2.6: Manhattan divided into four regions.

gion r. We choose to generate $f_r^{BA,k}(t)$ as the fraction of anticipated rides since we are interested in analyzing the change in the target number of drivers as the fraction of book-ahead rides increases.

As for the stochastic process $\{N_r^k(t) : t \in (kw, (k+1)w]\}$, at the beginning of each window k, we calibrate the demand rate λ_r^k corresponding to ride requests that will appear during the upcoming window in region r. In the following simulation, for simplicity, the demand rate varies across time-windows but is assumed constant within each time window; however, the proposed framework can be implemented using time-dependent demand rate functions by evaluating Equation 2.19. Moreover, even with window-constant demand rates, the Poisson distribution describing active drivers is time-varying within each window such that the mean is given by Equation 2.20. We emphasize that this transient analysis does not assume an equilibrium or steady-state conditions in any time window. The arrival rate for region 2 is shown in Figure 2.7; as observed, the demand rate increases rapidly showing the need for non-equilibrium methods. For the distribution $g_r^k(\cdot)$ representing ride duration, we use the empirical distribution that is derived from the observed rides in each region. Note that to analyze the change in the target number

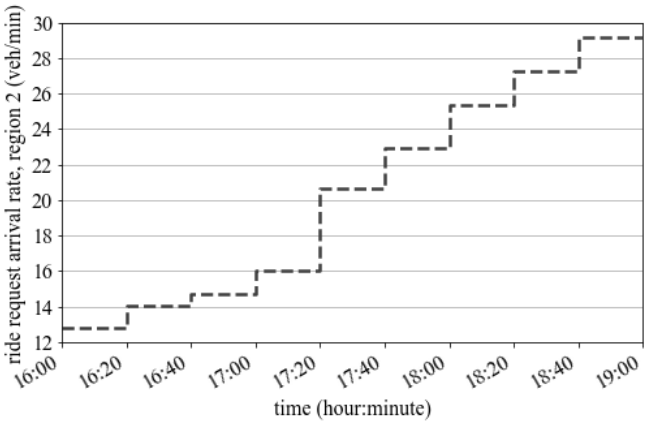

Figure 2.7: Arrival rate for ride requests that initiate in region 2.

of drivers with increasing book-ahead rides, we effectively assume that the arrival rate of non-reserved ride requests is $(1 - p_{BA})\lambda_r^k$ (where a fraction p_{BA} of the anticipated trips that will initiate during window k are book-ahead rides).

As illustrated in Figure 2.8, the proposed model for predicting the number of active rides (Section 2.2) accurately represents the observed data. In this Figure, for comparison with observed trip data, we consider that all rides are admitted and that there are no book-ahead rides (effectively assuming $N_r^k(t) = N_r^{k,\infty}(t)$). Recall that $N_r^k(t)$ represents the *predicted* non-reserved ride requests that will appear during window k; in contrast, during window $k+1$, the process $f_r^{P,k+1}(t)$ consists of observed trips (as given in the data) that differ from the previously predicted trips.

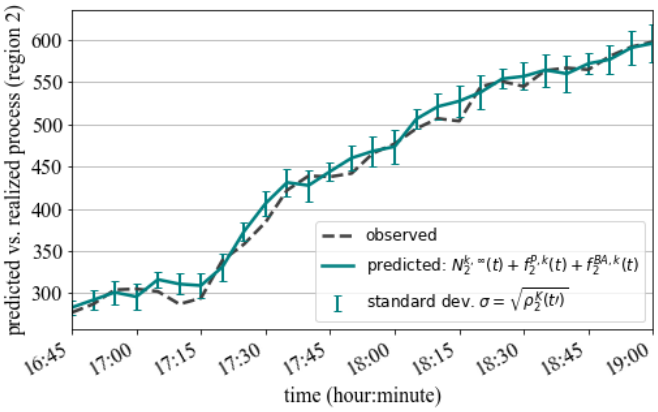

Figure 2.8: Predicted total number of active rides vs. observed number of active rides, where predictions were made over time windows with a duration of 20 minutes. The error bars correspond to one standard deviation of the time-dependent Poisson distribution characterizing $N_r^{k,\infty}$. In this figure, to compare with the observed trip data, we assume that all rides are admitted (i.e., we consider that $N_r^k(t) = N_r^{k,\infty}(t)$).

2.6.2 Upper bound on the blocking probability

To evaluate how tight is the upper bound in Inequality 2.22, we implement the admission control policy in region 2 and average the observed proportion of blocked rides B_r^k across time windows. For this upper bound numerical analysis, the assumptions involved in target evaluation and admission control apply; specifically, the total supply (active and idle) is maintained at the target level, drivers switch between active and idle within the region, and non-reserved rides are blocked if upon admission the total number of active rides would exceed the target at some point in time throughout the ride duration. Figure 9 shows the variation in the blocking proportion B_r^k relative to the upper bound δ. As observed, the blocking proportion B_r^k increases with larger tolerance values. We also observe that the blocking proportion increases with the fraction of book-ahead rides p_{BA}

as a result of fewer idle drivers being available for non-reserved rides.

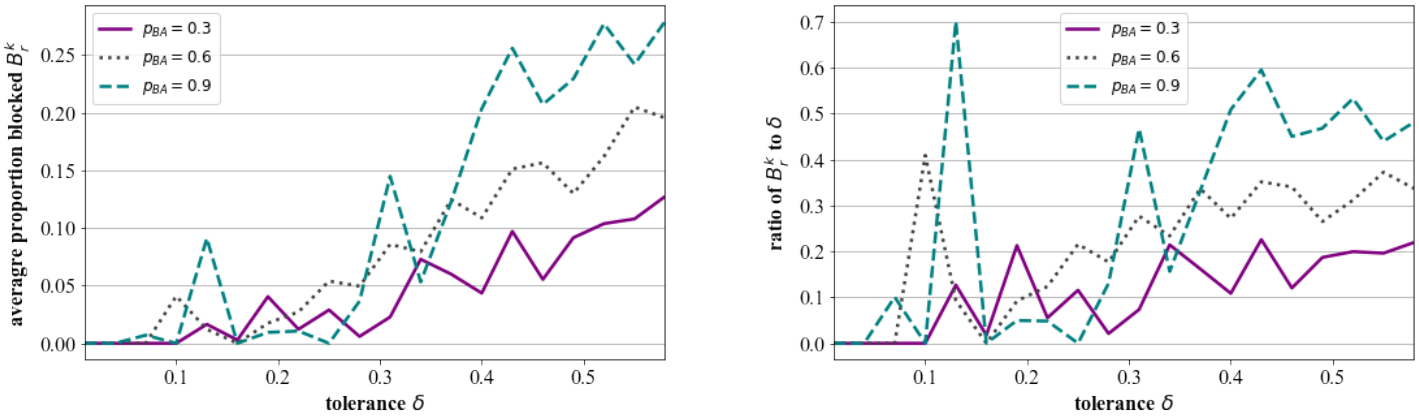

Figure 2.9: The change in observed blocking proportion B_r^k and the ratio B_r^k/δ relative to the upper bound δ.

2.6.3 Target computations, admission control, and minimum cost flow dispatching/rebalancing

Then, to account for the spatial distribution of demand and the variation in supply across regions, we implement the proposed framework in Sections 2.2–2.5 (see Figure 2.3). Different from Section 2.6.2, the demand moves between regions such that the supply deviates from the target, and we implement the min. cost flow program to maintain the target.

First, as mentioned in Section 2.6.1, we characterize the processes $\{f_r^{P,k}(t), f_r^{BA,k}(t), N_r^k(t) : t \in (kw, (k+1)w]\}$ representing the predicted number of active rides in each region r. Then, using the upper bound on the time-dependent blocking probability of the admission control policy, we determine the target number of drivers associated with every region r during the upcoming window. After that, at the beginning of the time window, we apply the driver dispatching/rebalancing mechanism

to attain the targets across regions. Then, throughout the time window, for every non-reserved ride request that is received, we implement the admission control policy to determine whether the request should be admitted or blocked; the received non-reserved ride requests are directly retrieved from the New York City data (as opposed to the predictions $N_r^k(t)$). We also implement the driver dispatching/rebalancing mechanism halfway through the time window. However, at the beginning of the time window we allow for total adjustments of the driver supply while halfway through the window we consider that only existing idle drivers can transition across adjacent regions. This process is then repeated for every time window.

For simulation purposes, we disregard the stochasticity of drivers entering and exiting the system across time windows. However, the admission control policy, target computations, and subsequent driver dispatching policy allow for a time-varying and stochastic variation in the supply that is joining or leaving the platform. In fact, target evaluation is based on the demand process and the admission control assumes that the target is maintained throughout the time window. Even if the actual supply deviates from the target, the admission control policy is still implemented by finding if there are any idle drivers and measuring the change in idle drivers relative to the target. On the other hand, the driver dispatching is only concerned with the instantaneous state of the supply relative to the target.

Note that the presented driver rebalancing strategy only uses information from the current time window. In other words, while the proposed state-dependent strategy does not assume steady-state conditions in a time-varying environment, it does not look into future windows to determine the current rebalancing recommendations. Alternative policies that predict future dynamics multiple windows in advance may also be effective since they would have more information on the anticipated variation in driver supply.

We apply the same framework for different fractions of book-ahead rides and

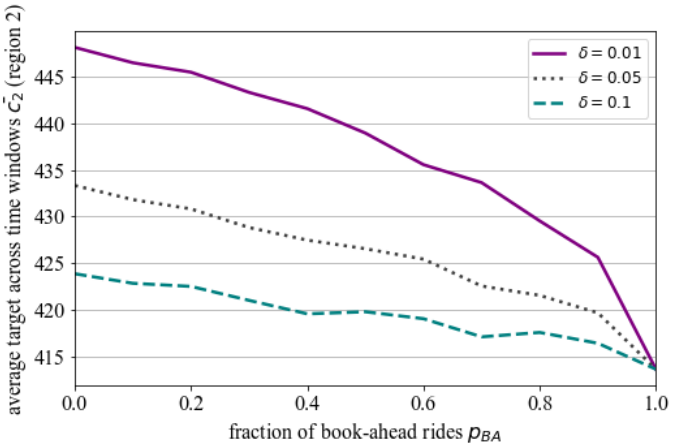

Figure 2.10: Change in the time-averaged target number of drivers with an increase in the fraction of book-ahead rides (for different quality of service thresholds δ). For each data point (i.e., every (p_{BA}, δ) pair), the plotted time-averaged target is the average of the corresponding value obtained from 30 different iterations of the proposed framework, where this averaging is needed due to the randomness in generation of the book-ahead profile $f_r^{BA,k}(t)$.

record the target c_r^k across windows. In Figure 2.10, we illustrate the change in targets for different fractions of book-ahead rides. In particular, we measure the time-averaged target $\bar{c}_r$ for increasing values of p_{BA} and different quality of service thresholds δ (as defined in Section 2.4.2, δ bounds the time-averaged blocking probability such that a lower value of δ indicates a higher quality of service). As expected, we observe that the target number of drivers increases with decreasing δ; this result implies that a larger number of drivers is needed to guarantee the reach time service requirement for a greater fraction of non-reserved ride requests. We also observe that the target number of drivers decreases as the fraction of book-ahead rides increases. The decrease in targets indicates that the number of drivers needed decreases with more information on anticipated trips.

For the simulation setting, the ratio of internal driver transitions $\sum_{(i,j)\in E} h_{ij}$ to the total flows $(\sum_{(i,j)\in E} h_{ij} + \sum_{i\in R} |h_i|)$ was approximately 0.5 when averaged across min-cost flow evaluations. The recommended external flows reflect the additional drivers needed

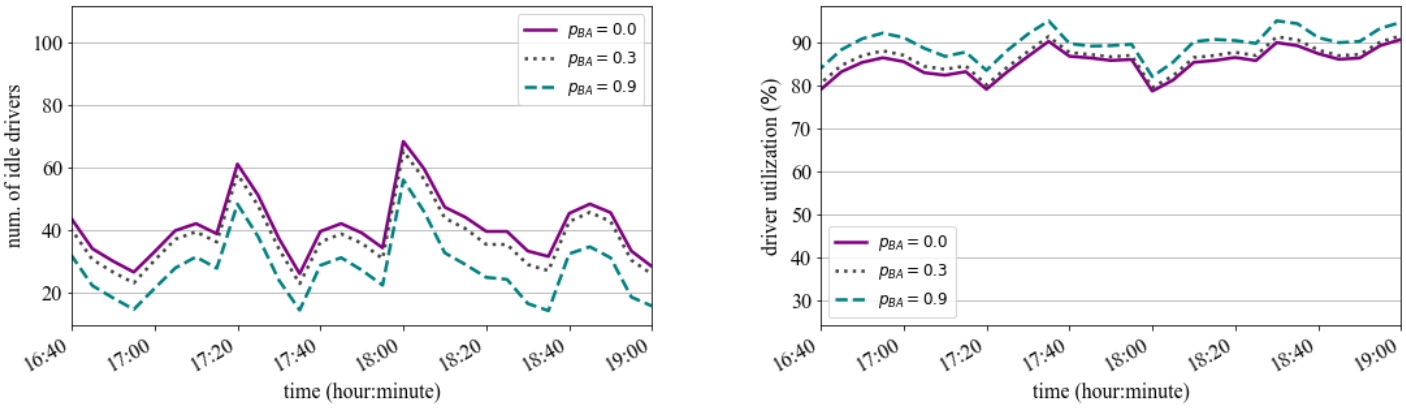

Figure 2.11: The number of idle drivers and the driver utilization rate 100*(active/(active+idle)) averaged across regions. The quality of service threshold δ is set at 0.01.

to satisfy increasing demand (Figure 2.7). This ratio depends on the demand rates, frequency of driver rebalancing, and the spatial distribution of regions. All these parameters would vary between different areas and time periods.

As the target decreases with increasing fractions of book-ahead rides, the number of idling drivers in the system also decreases. Figure 2.11 illustrates the average number of idling drivers for different reservation levels. We observe that when $p_{BA} = 0.9$ the average number of idle drivers can be up to 17.3 less than the corresponding value when $p_{BA} = 0.0$. This reduction in the number of idle drivers with increasing p_{BA} translates to a higher driver utilization rate.

Figure 2.12 illustrates the average number of rides that are blocked by the admission control policy (i.e., the reach time service requirement was not met for these rides). As shown, the average number of blocked rides increases with reservation levels. This increase in blocking results from the reduction in the overall number of drivers in the system. However, the fraction of blocked requests is (mostly) within the specified threshold $\delta = 0.01$. For $p_{BA} = 0.9$, the fraction of blocked requests slightly exceeds the level of ser-

vice threshold δ; this discrepancy may be attributed to the randomness in the system and the fact that the targets are not perfectly maintained throughout the entire time window.

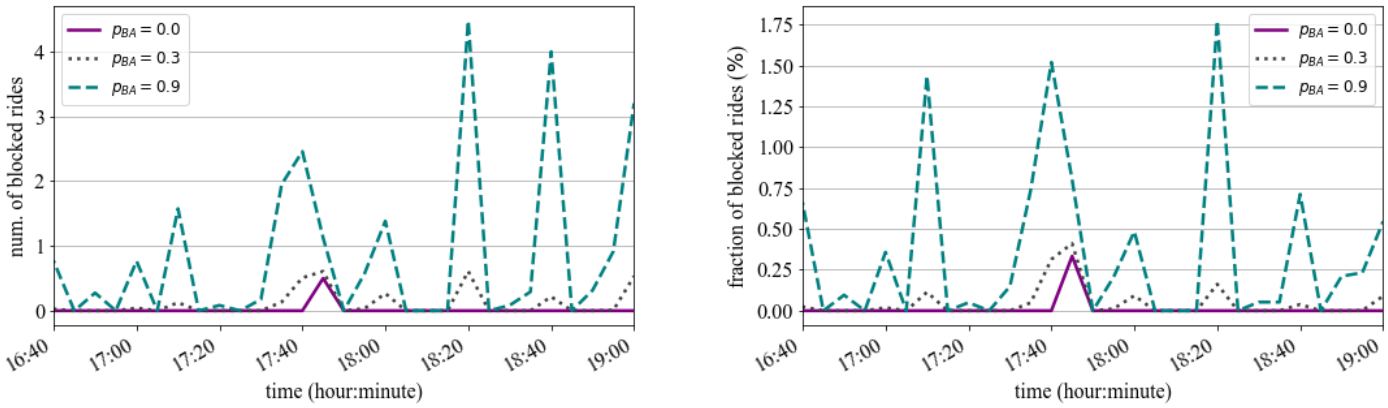

Figure 2.12: The number of blocked ride requests and the fraction of blocked requests 100*(blocked/(admitted+blocked)) averaged across regions. The quality of service threshold δ is set at 0.01.

The previous analysis assumed perfect compliance with inter-regional driver transitions at the simulation-specific driver rebalancing stages (beginning and mid-window). However, the drivers may not follow platform recommendations and that would result in greater difficulty maintaining the targets. Figure 2.13 shows the number of blocked rides and fraction of blocked rides in the worst-case scenario where drivers do not follow inter-regional transition recommendations. As observed, the number of blocked rides almost doubles in some cases and the fraction of blocked rides also increases up to 3.5%.

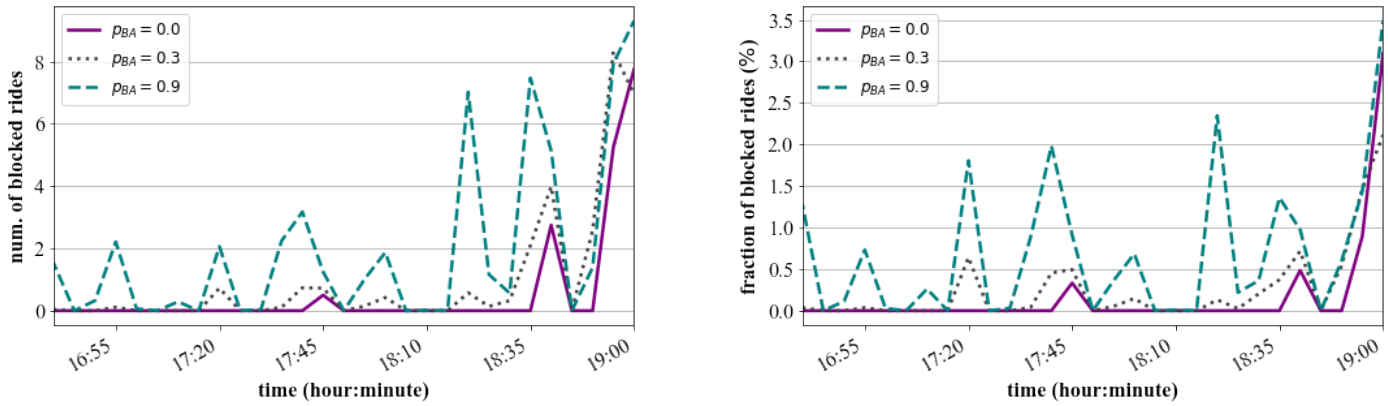

Figure 2.13: For the case when idle drivers do not follow platform-recommended transitions between regions, we observe an increase in the number blocked rides and the fraction of blocked rides. The quality of service threshold δ is set at 0.01.

2.7 Conclusion

In this chapter, we propose a model for transient analysis of stochasticity in ridesourcing systems. As opposed to steady-state equilibrium methods, we characterize the time-dependent state of the system and design control policies for managing driver supply. Furthermore, we incorporate book-ahead rides (reservations) in our framework and analyze the impact of book-ahead rides on driver supply management.

In more detail, we propose a state-dependent control policy that assigns drivers to observed ride requests with the objective of guaranteeing the reach time service requirement for book-ahead rides. Then, we derive a time-dependent upper bound on the performance of the control policy, where the performance of the policy is measured in terms of the probability of reach time service violations for non-reserved rides. Subsequently, this upper bound is used to determine the target number of drivers that probabilistically guarantees the reach time service requirement for non-reserved rides. The targets repre-

sent the total number of drivers that are associated with a region such that the drivers are either idling in the region or serving requests that initiate in the region. Then, considering a set of regions with different targets, we propose a driver dispatching/rebalancing optimization program that seeks to maintain the targets across regions. We show that the dispatching/rebalancing problem reduces to a minimum cost flow program that is solved on a transformed network.

The key findings are as follows: (1) For the desired reach time quality of service, an increase in the fraction of book-ahead rides leads to a reduction in the total number of drivers required. (2) This reduction in the total number of drivers is associated with a decrease in the number of idling drivers. (3) Once the driver supply is decreased, there is a greater risk that the reach time service requirement will be violated for anticipated non-reserved rides. However, the fraction of rides that experience increased reach time beyond the reach time service requirement is within a specified threshold, where this threshold dictates the target number of required drivers. (4) For Lyft rides in Manhattan, we observe rapid variations in demand rates that emphasize the need for transient analysis of ridesourcing dynamics.

The proposed model can be used for operation of ridesourcing systems. Specifically, the proposed control policy can be used for ensuring reach time priority for book-ahead rides, the target supply determines the number of drivers that would probabilistically guarantee the reach time service requirement for non-reserved rides, and the minimum cost flow program determines the necessary driver dispatching/rebalancing that is needed to maintain the targets.

More importantly, the proposed model can inform policy decisions that seek to maximize driver welfare and to reduce congestion externalities associated with ridesourcing platforms. In particular, for a given quality of service and reach time service requirement, policy makers can determine if the ridesourcing platform is employing an

excessive number of drivers by comparing the total number of drivers in the system to the target supply. In addition, our results suggest that policy makers should advocate for an increased fraction of book-ahead rides and supply management strategies that use this book-ahead information to reduce the number of idling drivers.

Chapter 3

Peak-Load Pricing and Demand Management for Ridesourcing Systems

3.1 Introduction

To limit the adverse impact of the supply-demand mismatch, platforms have also resorted to surge pricing as an effective tool for managing both supply and demand. During peak hours, surge pricing reduces the supply-demand mismatch by inhibiting passenger demand and at the same time attracting additional drivers to the surge location. However, surge pricing is controversial (Wang and Yang, 2019; Zuniga-Garcia et al., 2020). For example, drivers chasing the surge may reach the surge location after demand subsides while leaving passengers in other locations without service. To address such surge pricing drawbacks, we investigate alternative pricing policies where passengers in areas with high demand are offered the option to delay their trip in exchange for a reduced cost. In other words, we propose a pricing mechanism that induces users to travel during time periods when the predicted demand is low relative to the available supply (Yahia and Boyles, 2021).

Recent research on pricing in ridesourcing systems focuses on evaluating the optimal trip cost under supply-demand equilibrium (Bai et al., 2019; Banerjee et al., 2016; Bimpikis et al., 2019; Wang et al., 2016; Zha et al., 2018a,b), analyzing operational inefficiencies attributed to pricing (Xu et al., 2020; Zuniga-Garcia et al., 2020), and determining pricing strategies in transient (non-equilibrium) systems (Nourinejad and Ramezani, 2020). The majority of existing studies analyze equilibrium conditions within time periods where driver supply, passenger demand, or trip costs are time invariant. However, since supply and demand patterns vary rapidly across time, ridesourcing systems may

never attain equilibrium (Braverman et al., 2019). The proposed pricing strategy focuses on the transient nature of ridesourcing dynamics (Yahia et al., 2021b). We predict a time-dependent probabilistic characterization of anticipated ride requests. Then, when users request a ride, we use the demand predictions to compute the cost of each offered departure time alternative.

In particular, we consider that the platform dynamically provides users with multiple ride options, where each ride alternative consists of the trip cost at a delayed departure time. In turn, the passengers evaluate the utility of offered alternatives, and a multinomial logit model (MNL) is used to represent the probability that a passenger selects a specific alternative. To determine future demand peaks, we use a probabilistic characterization of anticipated spatiotemporal demand. Then, given the MNL model for passenger choice and the anticipated demand, we evaluate the trip cost for each offered departure time using an optimization problem that maximizes platform revenue subject to constraints that stagger demand peaks. The pricing policy is state-dependent, and it is successively implemented as ride requests appear across time.

The remainder of this chapter proceeds as follows: Section 3.2 presents the system model and the demand processes. Section 3.3 discusses departure time choice and its impact on anticipated demand. Section 3.4 discusses the platform pricing policy. Section 3.6 demonstrates the impact of the proposed pricing strategy using Lyft rides in Manhattan. Section 3.7 concludes the chapter.

3.2 System Model

The ridesourcing platform aims to price trip alternatives for ride requests that initiate in a geographic region $r \in \mathcal{R}$ (where $\mathcal{R}$ is the set of regions). As illustrated in Figure 3.1, we assume that the platform dynamically updates the offered alternatives

at the beginning of regular time intervals $\mathcal{U} = \{[u_l, u_{l+1}), [u_{l+1}, u_{l+2}), \ldots\}$. For example, at time u_l, the platform evaluates alternatives that will be offered to ride requests that will initiate during $[u_l, u_{l+1})$. Each alternative consists of a delayed departure time $\tau \in \mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$ within the time horizon $\mathsf{T}_{l+1} = [u_{l+1}, u_{l+1} + T]$ and an associated trip cost. The offered departure times $\tau \in \mathcal{T}$ do not have to coincide with end points of time intervals in $\mathcal{U}$.

Then, after the ride requests that initiate during $[u_l, u_{l+1})$ choose their trip departure time and cost, the platform generates a new set of alternatives (at time u_{l+1}) for ride requests that will initiate during $[u_{l+1}, u_{l+2})$. Similar to ride requests that previously initiated, the ride requests that initiate during $[u_{l+1}, u_{l+2})$ will be offered a new set of departure times $\tau \in \mathcal{T}'$ within the time horizon $\mathsf{T}_{l+2} = [u_{l+2}, u_{l+2} + T]$ and an associated trip cost for each departure time. The alternatives offered to ride requests that initiate during $[u_l, u_{l+1})$ are different from those offered to requests that initiate during $[u_{l+1}, u_{l+2})$, where this difference reflects variation of the system state between the time horizons T_{l+1} and T_{l+2}.

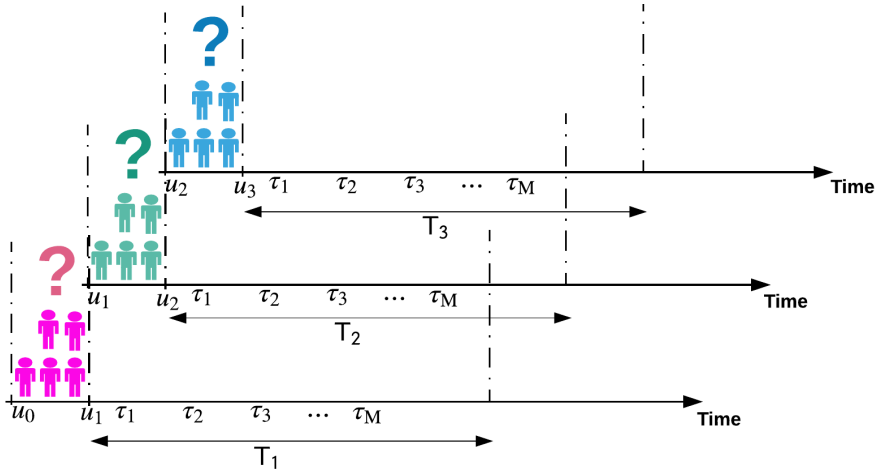

Figure 3.1: Time-dependent rolling horizon pricing mechanism.

Since the same pricing procedure is repeatedly used for ride requests that initiate

in any time interval $[u_l, u_{l+1}) \in \mathcal{U}$, we restrict our analysis to requests that initiate during $[u_0, u_1)$. For those rides, the offered departure time alternatives $\tau \in \mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$ are within the time horizon $\mathsf{T} = [u_1, u_1 + T]$.

In the subsequent analysis, we determine the trip cost of each departure time alternative $\tau \in \mathcal{T}$ based on the anticipated system state during T. We start by describing the predicted demand in Section 3.2.1. Then, in Section 3.2.2 we analyze the impact of the demand on the shortage in supply (change in idle drivers), and we define the *load* in a region as the process describing lost idle drivers. The resulting variation in idle drivers informs pricing strategies in Section 3.4.

In more detail, the *load process* that informs pricing decisions is derived from the anticipated trips that start or end in region r within T. We categorize those trips into *future* or *past* depending on whether the ride request is received prior to u_0 (past) or within T (future). Note that requests received prior to u_0 may start their trip within T due to users delaying their ride. Section 3.2.1 discusses past and future processes. Moreover, the users for which we are currently evaluating departure time alternatives (i.e., the users that will appear during $[u_0, u_1)$) are referred to as *now* users; Section 3.3 describes their choices and their impact on the load process.

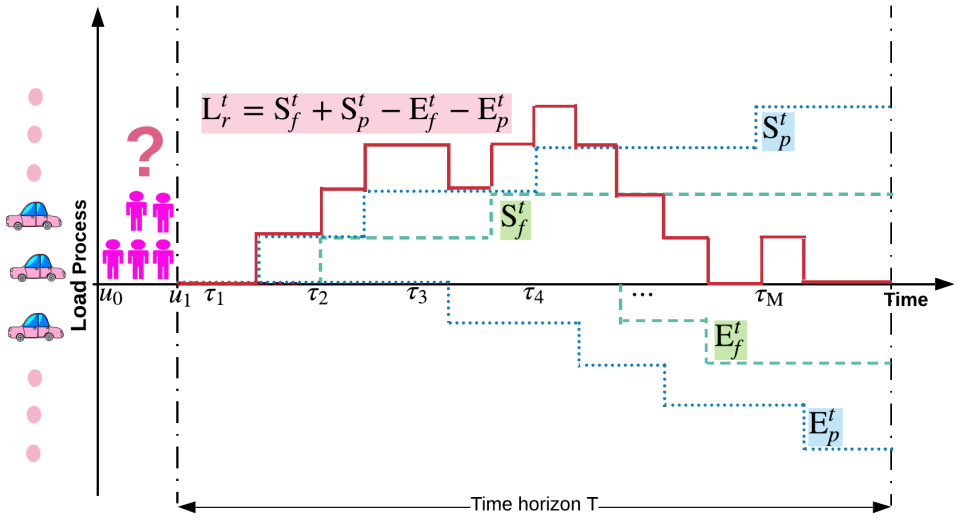

Figure 3.2: System model characterizing time-dependent ridesourcing dynamics in a region (zone) $r \in \mathcal{R}$. S_f^t represents the cumulative number of trips that start in r by time t and correspond to ride requests received in the *future* within T. E_f^t represents the cumulative number of trips that end in r by time t and correspond to ride requests received in the *future* within T. S_p^t represents the cumulative number of trips that start in r by time t and correspond to *past* ride requests that are received prior to u_0 (those rides start within T even though the request is received prior to u_0). E_p^t represents the cumulative number of trips that end in r by time t and correspond to *past* ride requests that are received prior to u_0 (those rides end within T). The load process is $L_r^t = S_f^t + S_p^t - E_f^t - E_p^t$.

3.2.1 Prediction of demand processes

We proceed by describing further the system state and the spatiotemporal demand during the time horizon T. The predicted demand processes dictate the supply-demand mismatch and the resulting shortage in idle drivers (high load).

As previously discussed, apart from *now* users, the demand during T has two components: (1) *past* demand that corresponds to ride requests received before u_0, and (2)

future demand that corresponds to ride requests that will be received during T. In the following demand characterization, we assume future ride requests do not delay their trip start time; this assumption ensures computational tractability and it is conservative in that it represents a worst case future demand scenario from the perspective of users that request a ride between $[u_0, u_1)$.

Future demand

First, we focus on future demand. For any pair of regions $i, j \in \mathcal{R}$, we consider that future ride requests for users traveling between i and j will appear according to a Poisson process. In addition, we assume that the platform can estimate the rate of ride requests $\{\lambda_{ij}^t : t \in \mathsf{T}\}$. For simplicity, we consider that the rate λ_{ij} is fixed within the horizon T; however, the proposed mechanism can be easily generalized to cases with a time-dependent ride request rate. We also assume that the ride duration will be generally distributed such that the service time distribution for rides between i, j is denoted by $g_{ij}(\cdot)$ and its cumulative density function is $G_{ij}(\cdot)$.

Thus, at time u_0 and for any region $r \in \mathcal{R}$, the platform can characterize two different predicted demand process $\{\mathrm{S}_f^t, \mathrm{E}_f^t : t \in \mathsf{T}\}$ associated with *future* ride requests. These processes are stochastic since they are determined by ride requests that appear according to a Poisson process and ride durations that are generally distributed. Moreover, these processes depend on the spatial distribution of demand across the regions in $\mathcal{R}$. The process S_f^t represents the *cumulative* number of future rides that will start in region $r \in \mathcal{R}$ by time $t \in \mathsf{T}$. A ride *starts* when the driver is assigned to fulfill the trip (i.e., the driver is no longer idle). On the other hand, E_f^t represents the *cumulative* number of rides that end in region $r \in \mathcal{R}$ by time $t \in \mathsf{T}$ (once a trip ends, region r would gain an idle driver). The processes $\{\mathrm{S}_f^t, \mathrm{E}_f^t : t \in \mathsf{T}\}$ are illustrated in Figure 3.2.

We assume that the processes $\{\mathrm{S}_f^t, \mathrm{E}_f^t : t \in \mathsf{T}\}$ are unbounded. An equivalent

assumption is that all ride requests can be served. Thus, we may observe that the predicted number of trips that start in r is greater than the number of available idle drivers throughout the time horizon; in practice, this would correspond to distant drivers from an external region being dispatched to serve requests that start in r. In other words, the demand processes corresponds to trips starting in r or trips ending in r (even if the driver had to be dispatched from an external region to serve requests in r).

Previously observed demand

In addition to the future demand, we assume that the platform can accurately determine the trip start time and duration for *previously* observed ride requests (i.e., the platform has full information on ride requests received prior to time u_0). Thus, for each region $r \in \mathcal{R}$, the platform can characterize *deterministic* processes $\{\mathrm{S}_p^t, \mathrm{E}_p^t : t \in \mathsf{T}\}$ corresponding to the cumulative number of starts/ends that occur during T given that the request was received prior to time u_0. S_p^t represents prior ride requests that start in region r by time $t \in \mathsf{T}$. Similarly, E_p^t represents prior ride requests that end in region r by time $t \in \mathsf{T}$. We emphasize that $\{\mathrm{S}_p^t, \mathrm{E}_p^t : t \in \mathsf{T}\}$ are restricted to starts or ends that occur within T and that requests received prior to u_0 may start their trip within T due to users delaying their ride.

3.2.2 Predicted load process

Given these demand processes, we can define the *load process* L_r^t, where L_r^t corresponds to the change in the number of idle drivers between u_1 and $t \in \mathsf{T}$. In particular, we can express L_r^t in Equation 3.1 as the number of trips ending in r subtracted from the number of trips that start in r. Observe that if the trips starting in region r is greater than the trips ending in region r the load will increase; thus, large load values indicate that

there is a net decrease in idle drivers. Note that L_r^t is independent of the users that appear between $[u_0, u_1)$. In other words, L_r^t is either caused by future demand or prior demand such that it is independent of the users we seek to price. The load process L_r^t is illustrated in Figure 3.2.

$$L_r^t = S_f^t + S_p^t - E_f^t - E_p^t \tag{3.1}$$

In Section 3.4, we will use the expected value of L_r^t to compute the price of each offered departure time alternative. The pricing strategy aims to disperse users that initiate between $[u_0, u_1)$ away from periods with high expected load $\mathbb{E}\left[L_r^t\right]$. Thus, we proceed by evaluating $\mathbb{E}\left[L_r^t\right]$ given in Equation 3.2.

$$\mathbb{E}\left[L_r^t\right] = \mathbb{E}\left[S_f^t\right] + S_p^t - \mathbb{E}\left[E_f^t\right] - E_p^t \tag{3.2}$$

Expected number of future ride requests that start in region $r \in \mathcal{R}$

We start by deriving $\mathbb{E}\left[S_f^t\right]$. Since future ride requests are received according to a Poisson process, the expected number of trips starting in r by time t is given in Equation 3.3. Observe that $\mathbb{E}\left[S_f^t\right]$ is time-dependent indicating lost idle drivers as time progresses.

$$\mathbb{E}\left[S_f^t\right] = \sum_{j \in \mathcal{R}} \lambda_{rj}(t - u_1) \tag{3.3}$$

Expected number of future ride requests that end in region $r \in \mathcal{R}$

Next, we derive $\mathbb{E}\left[E_f^t\right]$, the expected number of future ride requests that end in r by time t. Recall that we assume all requests could be served. In addition, for demand traveling from $j \in \mathcal{R}$ to r, we assume that future ride requests will be received according to a Poisson process with rate λ_{jr} and that the ride duration has a general distribution

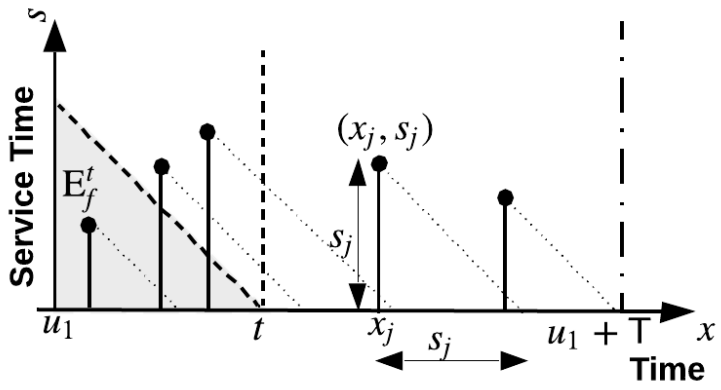

Figure 3.3: Service time vs. arrival time for future rides that are received after time u_1. The dotted diagonal lines represent the decrease in remaining service time as the user is being served. For any time t, the number of users that have completed service is the number of points in the shaded area. For all such points, the intersection of the associated dotted diagonal line with the x-axis is less than t. The shaded area also corresponds to users that are served by time t in a transient M/GI/∞ queue that starts empty at time u_1.

$g_{jr}(\cdot)$ with the CDF being $G_{jr}(\cdot)$. In the following, we use a graphical approach to show that E_f^t has a time-dependent Poisson distribution and $\mathbb{E}\left[\mathrm{E}_f^t\right]$ is its time-dependent mean (Prékopa, 1958).

In Figure 3.3, let x_j denote the trip start time (ride request initiation) of the j^{th} future user that appears according to the Poisson process. Note that the trip may start in an external region (provided it starts during T as a future ride request). Moreover, let s_j denote the corresponding service time for the j^{th} future user. As shown in Figure 3.3, we observe that (x_j, s_j) is a random point in the two-dimensional plane $[u_1, u_1 + T] \times [0, \infty)$ that represents the trip start time and service duration. Thus, for any two-dimensional set S in $[u_1, u_1 + T] \times [0, \infty)$, the number of points in the set represents random sampling of the ride requests Poisson process; therefore, the number of points in the set S is *Poisson distributed*. We also know that disjoint two-dimensional sets correspond to independent sampling of a Poisson process; this implies that the number of points in each set is independent of other disjoint sets.

Furthermore, if we isolate an infinitesimal two-dimensional square set with an area $ds(dx)$, we can see that the mean number of points in that set is $\lambda_{jr}dx\left(g_{jr}(s)(ds)\right)$. Thus, for any two-dimensional set S, the intensity of the two-dimensional Poisson distribution is $\lambda_{jr}g_{jr}(s)$. In other words, the distribution of points defined as (arrival time, service duration) is Poisson over the two-dimensional space, and the *expected* number of points for any set S is given by $\int_S \lambda_{jr}g_{jr}(s)dsdx$.

As a result, to determine the *expected* number of arrivals from region j, we evaluate the integral $\int_S \lambda_{jr}g_{jr}(s)dsdx$ over the shaded area illustrated in Figure 3.3. This shaded area represents trips that started in j and have completed in region r prior to time $t \in \mathsf{T}$. In particular, for each (arrival time, service time) pair associated with a specific user, the diagonal line represents the decrease in remaining service time as the user is being served. Note that for all users in the shaded area, the corresponding diagonal line intersects the x-axis prior to time t; this indicates that the trip terminates in region r before time t.

In addition, observe that evaluating the integral $\int_S \lambda_{jr}g_{jr}(s)dsdx$ over the shaded area is equivalent to calculating the expected number of served users in a transient M/GI/∞ queue that starts empty at u_1, where the M/GI/∞ queue has an arrival rate λ_{jr} and a general service distribution g_{jr} (the queue has infinite servers since all users can be served).

Then, to compute $\mathbb{E}\left[\mathrm{E}_f^t\right]$, the expected number of total trips that start in *any region* and end in r by time t, we sum the integral $\int_S \lambda_{jr}g_{jr}(s)dsdx$ across all regions $j \in \mathcal{R}$ (where the integral is evaluated using the bounds of the shaded area). The resulting expression for $\mathbb{E}\left[\mathrm{E}_f^t\right]$ is given in Equation 3.4. Similar to $\mathbb{E}\left[\mathrm{S}_f^t\right]$, we observe that $\mathbb{E}\left[\mathrm{E}_f^t\right]$ is time-dependent indicating the change in load across time.

$$\mathbb{E}\left[\mathrm{E}_f^t\right] = \sum_{j\in\mathcal{R}} \int_{u_1}^{t} \int_{0}^{t-x} \lambda_{jr}g_{jr}(s)dsdx = \sum_{j\in\mathcal{R}} \lambda_{jr} \int_{0}^{t-u_1} G_{jr}(x)dx \tag{3.4}$$

3.3 Passenger Price and Departure Time Choice

Now that the load process in Equation 3.2 can be evaluated from past and future trips (Equations 3.3 and 3.4 for future starts/ends). We proceed to analyze the choices of *now* users and their impact on the total trip starts/ends. Section 3.3.1 discusses the multinomial logit model representing user choice, and Section 3.3.2 discusses the impact of choice probabilities on trip starts and ends. Then, in Section 3.4, we use the passenger choices and their impact on future supply-demand to determine the price of each offered departure time alternative.

3.3.1 The multinomial logit model

The probability $p_k(c_k, \tau_k)$ of a passenger choosing a particular departure time alternative $\tau \in \mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$ is determined by random utility theory. In particular, we use a multinomial logit (MNL) model.

The MNL model and the subsequent pricing optimization problem in Section 3.4 use the time-dependent surcharge c_k instead of the full trip cost, where the full trip cost consists of the surcharge added to a base fare. Since the base fare for each user is time-invariant (depending on factors such as the length of the trip, operational costs, type of service etc.), it does not factor into the departure time choice. In other words, the base fare would be the same for different departure time alternatives and the difference in cost is determined solely by the time-dependent surcharge. Moreover, while the base fare differs across users, the same surcharge is assigned for all users that choose a specific departure time alternative.

A MNL for traveler choice can be specified by defining the utility V of travel. Assume that the surcharge c_k is less than c_1 such that travelers would only delay their trip for a reduced trip fare. Let $a_k = c_1 - c_k$ be the *savings* that result from departing at time

$\tau_k \in \mathcal{T}$, and let $d_k = \tau_k - \tau_1$ be the associated delay. In addition, define $V_k = \beta_c a_k + \beta_d d_k$ to be the utility of choosing a specific departure time $\tau_k \in \mathcal{T}$. The resulting utility of departing now, at τ_1, would be $V_1 = 0$.

Since a_k is restricted to be greater than zero and it represents savings, it is expected that β_c is positive (increased utility with greater savings). The parameter β_d represents the sensitivity towards delay and it is expected to be negative. It is assumed that the service provider can estimate those parameters from past data.

Then, the MNL probabilities $p_k(c_k, \tau_k)$ are given in Equations 3.5 and 3.6. For brevity, we denote $p_k(c_k, \tau_k)$ as p_k.

$$p_1 = \frac{1}{1 + \sum_{\tau_j \in \mathcal{T} \setminus \{\tau_1\}} e^{\beta_c a_j + \beta_d d_j}} \tag{3.5}$$

$$p_k = \frac{e^{\beta_c a_k + \beta_d d_k}}{1 + \sum_{\tau_j \in \mathcal{T} \setminus \{\tau_1\}} e^{\beta_c a_j + \beta_d d_j}} \qquad \forall \tau \in \mathcal{T} \setminus \{\tau_1\} \tag{3.6}$$

The MNL model is not particularly suitable for departure time choice due to the independence of irrelevant alternatives (IIA) property. With an MNL model, departure times that are adjacent to each other do not exhibit increased sensitivity compared to non-adjacent ones. However, there might be excluded exogenous factors that cause correlations among adjacent time slots. Alternative models such as the ordered generalized extreme-value (OGEV) model are more appropriate; those models place adjacent departure times within nests (Bhat, 1998; Small, 1987, 1994). Refer to Train (2009) for more on MNL assumptions and the general extreme-value family of models which includes the MNL as a special case of nested variations.

That said, MNL models are readily available within the toolkit of metropolitan planning organizations and have been previously used for departure time choice analysis (Saleh and Farrell, 2005; Steed and Bhat, 2000). In fact, Steed and Bhat (2000) state

that an MNL model was adequate for departure time choice in terms of data fit and that attempts at estimating an OGEV resulted in estimates that are inconsistent with utility maximization theory (logsum parameter exceeding 1). Small (1987) also reported violation of random utility maximization principles in attempts at estimating OGEV models.

Moreover, for the optimization program in section 3.4, initial attempts at using an OGEV or nested models instead of an MNL led to non-convex optimization programs that can not be reduced into convex equivalents. The primary difficulty arose from the inclusion of log-sum coefficients.

3.3.2 Impact of choices on the load process

Given the MNL probability p_k that users arriving *now* between $[u_0, u_1)$ select to depart at time τ_k, we determine the impact of these choices on the load process. Similar to the analysis approach of future rides in Section 3.2, we determine the number of trips that start/end in T given that the ride request will be received during $[u_0, u_1)$ and the departure time choice follows from the MNL model.

In more detail, the additional load δ_r^t at time $t \in \mathsf{T}$ that is associated with users that appear between $[u_0, u_1)$ is shown in Equation 3.7. The term S_n^t represents the cumulative number of trips that start by time t and correspond to users requesting service between $[u_0, u_1)$. Similarly, the term E_n^t represents the cumulative number of trips that end by time t and correspond to users requesting service between $[u_0, u_1)$. The expected additional load $\mathbb{E}\left[\delta_r^t\right]$ is given in Equation 3.8. In what follows, we evaluate $\mathbb{E}\left[\delta_r^t\right]$.

$$\delta_r^t = \mathrm{S}_n^t - \mathrm{E}_n^t \tag{3.7}$$

$$\mathbb{E}\left[\delta_r^t\right] = \mathbb{E}\left[\mathrm{S}_n^t\right] - \mathbb{E}\left[\mathrm{E}_n^t\right] \tag{3.8}$$

Note that the departure time alternatives and their prices are generated at time u_0. Thus, the platform needs to characterize the anticipated arrival rate and trip duration for requests that appear in $[u_0, u_1)$. Similar to the future ride requests, the platform would estimate the arrival rate λ_{rj} for trips between regions $r, j \in \mathcal{R}$. The ride duration also follows a general distribution $g_{rj}(\cdot)$ with CDF $G_{rj}(\cdot)$.

Expected number of ride requests that start in region $r \in \mathcal{R}$ for users requesting service between $[u_0, u_1)$

Then, we derive $\mathbb{E}\left[\mathrm{S}_n^t\right]$, the expected cumulative number of trips that start before time t in r and are associated with requests that will be received during $[u_0, u_1)$. The expression for $\mathbb{E}\left[\mathrm{S}_n^t\right]$ is given in Equation 3.9; we arrive at this expression by calculating the *total* expected number of requests received between $[u_0, u_1)$ and multiplying by the probability that those requests choose to depart prior to time $t \in \mathsf{T}$ (i.e., they choose a departure time $\tau \in \mathcal{T} = \{\tau_1, \tau_2, ..\tau_M\}$ that is less than t).

$$\mathbb{E}\left[\mathrm{S}_n^t\right] = \left[\sum_{j \in \mathcal{R}} \lambda_{rj}(u_1 - u_0)\right] \sum_{\tau_k \in \mathcal{T} : \tau_k \leq t} p_k \tag{3.9}$$

Expected number of ride requests that end in region $r \in \mathcal{R}$ for users requesting service between $[u_0, u_1)$

Similarly, we derive $\mathbb{E}\left[\mathrm{E}_n^t\right]$, the expected number of trips that end by time t in r and are associated with requests that will be received during $[u_0, u_1)$. The expression for $\mathbb{E}\left[\mathrm{E}_n^t\right]$ is given in Equation 3.10. To obtain Equation 3.10, we multiply the expected *total* number of users that would be received between $[u_0, u_1)$ and are destined to r by the probability that their trip ends before time $t \in \mathsf{T}$; in turn, the probability that the trip ends before time t is the product of the probability that the trip starts prior to time t and the

probability that the ride duration is less than the difference between t and the start time.

$$\mathbb{E}\left[\mathrm{E}_n^t\right] = \lambda_{rr}(u_1 - u_0) \sum_{\tau_k \in \mathcal{T}: \tau_k \leq t} p_k G_{rr}(t - \tau_k) \tag{3.10}$$

3.4 Peak-Load Pricing

After defining the load process and the impact of *now* users on the trip start/ends, we are able to identify time periods where the load is high and we would want the probability of departure at that time to be low. In this section, we describe the platform pricing strategy that evaluates the optimal trip costs. Recall that the trip costs determine the departure time probabilities via the MNL model; in turn, departure time probabilities determine the impact of now users on trip starts/ends (see Equations 3.9 and 3.10). Thus, the optimal costs are those that maximize revenue while ensuring that the probability of departure during peak-load periods is limited. We seek to find the trip cost associated with each departure time alternative $\tau \in \mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_M\}$ offered to users that initiate between $[u_0, u_1)$.

3.4.1 Platform revenue maximization

The platform optimization problem is shown in formulation 3.11–3.16. The term $\sum_{\tau_k \in \mathcal{T}} c_k p_k$ in the maximization objective refers to the expected revenue *per ride*; it is the sum across alternatives of the surcharge multiplied by the choice probability.

In addition to revenue maximization, the pricing strategy also aims to restrict the load process, where an increase in the load process indicates lost idle drivers. To that end, the component of the objective given by wz along with constraints 3.12 and 3.16 minimize the *increase* in load across departure time alternatives. Note that constraint 3.12 is linear in the departure time probabilities. The term w is a constant that weights the two

components of the objective.

Constraints 3.13 and 3.14 represent the MNL model that relates the surcharge to the choice probabilities. Constraint 3.15 guarantees that the savings are positive (i.e., that the future departure time alternatives have a lower surcharge). Thus, formulation 3.11–3.16 finds the optimal surcharges that maximize platform revenue, minimize peaks in the load process, and reflect user choices.

$$\max_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k,c_k:\tau_k\in\mathcal{T}\setminus\{\tau_1\},z}} \quad \sum_{\tau_k\in\mathcal{T}} c_k p_k - wz \tag{3.11}$$

$$\text{s.t.} \quad \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_{k+1}}\right] + \mathbb{E}\left[\delta_r^{\tau_{k+1}}\right]\right) - \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_k}\right] + \mathbb{E}\left[\delta_r^{\tau_k}\right]\right) \leq z \qquad \forall \tau_k \in \mathcal{T}\setminus\{\tau_M\} \tag{3.12}$$

$$p_1 = \frac{1}{1+\sum_{\tau_j\in\mathcal{T}\setminus\{\tau_1\}} e^{\beta_c a_j+\beta_d d_j}} \tag{3.13}$$

$$p_k = \frac{e^{\beta_c a_k+\beta_d d_k}}{1+\sum_{\tau_j\in\mathcal{T}\setminus\{\tau_1\}} e^{\beta_c a_j+\beta_d d_j}} \qquad \forall \tau \in \mathcal{T}\setminus\{\tau_1\} \tag{3.14}$$

$$a_k \geq 0 \qquad \forall \tau \in \mathcal{T}\setminus\{\tau_1\} \tag{3.15}$$

$$z \geq 0 \tag{3.16}$$

The optimization problem in 3.11–3.16 is non-convex and it has a nonlinear objective function with nonlinear constraints. Specifically, the choice probabilities p_k are a nonlinear function of the surcharge decision variables c_k. In Section 3.4.2, we reformulate the optimization problem to arrive at a convex program. First, we reduce formulation 3.11–3.16 to an equivalent formulation where the decision variables are only p_k and z. Then, we show that (in terms of p_k and z) the objective is convex and the constraints form a convex set.

Observe that the utilities only depend on the *difference* between c_1 and c_k (for some $\tau_k \in \mathcal{T}\setminus\{\tau_1\}$). If both c_1 and c_k are decision variables, then an infinite possible com-

binations of those variables would result in the same set of utilities. The optimization problem is then unbounded. To be precise, since many combinations of costs lead to the same probabilities, as long as the problem is feasible (demand shaving constraints are satisfied), infinitely high values of the cost variables would be chosen to maximize the objective. This is especially problematic since the option of not choosing any departure time is not considered. Thus, we assume that c_1 is a fixed constant, and the costs c_k at $\tau_k \in \mathcal{T}\backslash\{\tau_1\}$ are decision variables, where the constraints impose that $c_k \leq c_1$.

3.4.2 Convex revenue maximization formulation given passenger choice

To reformulate the problem into a convex program, we start by replacing the maximization in 3.11–3.16 with the *minimization* in 3.17–3.22. This can be thought of as replacing the maximization of revenue objective with one that minimizes losses. Those two forms are equivalent. As shown below, the revised objective is in terms of a_k.

The revised formulation is as follows:

$$\min_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k:\tau_k\in\mathcal{T}\backslash\{\tau_1\},z}} \quad \sum_{\tau_k\in\mathcal{T}\backslash\{\tau_1\}} a_k p_k + wz \tag{3.17}$$

$$\text{s.t.} \quad \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_{k+1}}\right] + \mathbb{E}\left[\delta_r^{\tau_{k+1}}\right]\right) - \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_k}\right] + \mathbb{E}\left[\delta_r^{\tau_k}\right]\right) \leq z \qquad \forall \tau_k \in \mathcal{T}\backslash\{\tau_M\} \tag{3.18}$$

$$p_1 = \frac{1}{1 + \sum_{\tau_j\in\mathcal{T}\backslash\{\tau_1\}} e^{\beta_c a_j + \beta_d d_j}} \tag{3.19}$$

$$p_k = \frac{e^{\beta_c a_k + \beta_d d_k}}{1 + \sum_{\tau_j\in\mathcal{T}\backslash\{\tau_1\}} e^{\beta_c a_j + \beta_d d_j}} \qquad \forall \tau \in \mathcal{T}\backslash\{\tau_1\} \tag{3.20}$$

$$a_k \geq 0 \qquad \forall \tau \in \mathcal{T}\backslash\{\tau_1\} \tag{3.21}$$

$$z \geq 0 \tag{3.22}$$

Claim. *Solving for an optimal solution to the maximization formulation 3.11–3.16 is equivalent to solving for the minimum of formulation 3.17–3.22*

Proof. We show that the objectives of the two formulations are equivalent as follows:

$$\max_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k,c_k:\tau_k\in\mathcal{T}\setminus\{\tau_1\},z}} \quad \sum_{\tau_k\in\mathcal{T}} c_k p_k - wz \tag{3.23}$$

$$\Leftrightarrow \max_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k:\tau_k\in\mathcal{T}\setminus\{\tau_1\},z}} \quad \sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} (c_1 - a_k)\, p_k + c_1 p_1 - wz \tag{3.24}$$

$$\Leftrightarrow \max_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k:\tau_k\in\mathcal{T}\setminus\{\tau_1\},z}} \quad -\sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} a_k p_k + c_1 \cancelto{1}{\sum_{\tau_k\in\mathcal{T}} p_k} - wz \tag{3.25}$$

$$\Leftrightarrow \min_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k:\tau_k\in\mathcal{T}\setminus\{\tau_1\},z}} \quad \sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} a_k p_k + wz \tag{3.26}$$

□

The revised formulation is still a non-convex optimization problem since p_k is a nonlinear function of a_k. Thus, in the subsequent reformulation 3.27–3.32, we reduce the optimization problem 3.17–3.22 into a *convex program* in terms of p_k and z. In more detail, the MNL constraints in 3.19 and 3.20 implicitly force the probabilities to sum to one and to be between zero and one. In what follows, since the problem is solved in terms of p_k, the constraints on the probabilities are explicitly stated and the MNL constraints are dropped. The objective 3.17 reduces to the convex function 3.27. Constraint 3.21 can be rewritten as constraint 3.31 which is linear in p_k.

The revised formulation is as follows:

$$\min_{p_k:\tau_k\in\mathcal{T},z} \quad \frac{1}{\beta_c}\left[\sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} p_k \log(p_k) - \beta_d d_k p_k\right] - \frac{1}{\beta_c}(1-p_1)\log(p_1) + wz \tag{3.27}$$

$$\text{s.t.} \quad \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_{k+1}}\right] + \mathbb{E}\left[\delta_r^{\tau_{k+1}}\right]\right) - \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_k}\right] + \mathbb{E}\left[\delta_r^{\tau_k}\right]\right) \le z \qquad \forall \tau_k \in \mathcal{T}\setminus\{\tau_M\} \tag{3.28}$$

$$\sum_{\tau_k \in \mathcal{T}} p_k = 1 \tag{3.29}$$

$$0 \leq p_k \leq 1 \qquad \forall \tau \in \mathcal{T} \tag{3.30}$$

$$p_k \geq e^{\beta_d d_k} p_1 \qquad \forall \tau \in \mathcal{T} \backslash \{\tau_1\} \tag{3.31}$$

$$z \geq 0 \tag{3.32}$$

Claim. *Solving for an optimal solution to 3.17–3.22 is equivalent to solving for the minimum of formulation 3.27–3.32*

Proof. First, we show that the two objectives are equivalent.

From Equation 3.19, we know that $\log(p_1) = -\log\left(1 + \sum_{\tau_j \in \mathcal{T} \backslash \{\tau_1\}} e^{\beta_c a_j + \beta_d d_j}\right)$

From Equation 3.20, we know that $\log(p_k) = \beta_c a_k + \beta_d d_k - \log\left(1 + \sum_{\tau_j \in \mathcal{T} \backslash \{\tau_1\}} e^{\beta_c a_j + \beta_d d_j}\right)$

Thus, $\log(p_k) = \beta_c a_k + \beta_d d_k + \log(p_1)$

Rearranging, we can write a_k as follows:

$$a_k = \frac{1}{\beta_c} \left[\log(p_k) - \log(p_1) - \beta_d d_k\right] \tag{3.33}$$

Thus, $a_k p_k = \frac{1}{\beta_c} \left[p_k \log(p_k) - p_k \log(p_1) - \beta_d d_k p_k\right]$

Then,

$$\sum_{\tau_k \in \mathcal{T} \backslash \{\tau_1\}} a_k p_k = \frac{1}{\beta_c} \sum_{\tau_k \in \mathcal{T} \backslash \{\tau_1\}} \left[p_k \log(p_k) - \beta_d d_k p_k\right] - \frac{1}{\beta_c} \log(p_1) \sum_{\tau_k \in \mathcal{T} \backslash \{\tau_1\}} p_k \tag{3.34}$$

($\sum_{\tau_k \in \mathcal{T} \backslash \{\tau_1\}} p_k = 1 - p_1$)

$$= \frac{1}{\beta_c} \sum_{\tau_k \in \mathcal{T} \backslash \{\tau_1\}} \left[p_k \log(p_k) - \beta_d d_k p_k\right] - \frac{1}{\beta_c} (1 - p_1) \log(p_1) \tag{3.35}$$

where equation 3.34 follows from the requirement that the probabilities sum to one.

This implies that:

$$\min_{\substack{p_k:\tau_k\in\mathcal{T},\\ a_k:\tau_k\in\mathcal{T}\setminus\{\tau_1\},z}} \quad \sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} a_k p_k + wz$$

is equivalent to:

$$\min_{p_k:\tau_k\in\mathcal{T},z} \quad \frac{1}{\beta_c}\left[\sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} p_k \log(p_k) - \beta_d d_k p_k\right] - \frac{1}{\beta_c}(1-p_1)\log(p_1) + wz$$

Furthermore, we show that constraint 3.21 can be expressed in terms of p_k as follows:
From Equation 3.33, since β_c is positive (sensitivity to savings), we know that $a_k \geq 0$ if $\log(p_k) - \log(p_1) - \beta_d d_k \geq 0$.
Thus, $a_k \geq 0$ if $\log(\frac{p_k}{p_1}) \geq \log\left(e^{\beta_d d_k}\right)$, and this implies that $a_k \geq 0$ if $p_k \geq e^{\beta_d d_k} p_1$ □

Thus, since the constraints form a convex set in p_k and z, we can show that the formulation 3.27–3.32 is a convex program by verifying that the objective is a convex function. In this case, we show that the Hessian matrix associated with the objective is positive semidefinite.

Claim. *The objective function*
$\mathrm{F} = \frac{1}{\beta_c}\left[\sum_{\tau_k\in\mathcal{T}\setminus\{\tau_1\}} p_k \log(p_k) - \beta_d d_k p_k\right] - \frac{1}{\beta_c}(1-p_1)\log(p_1) + wz$
is convex

Proof. Observe that the objective is separable with respect to each decision variable $\{p_k : \tau_k \in \mathcal{T}\}, z$
Then, we can easily determine the second derivative with respect to each variable and construct the corresponding diagonal Hessian matrix as follows:
$\frac{\partial^2 \mathrm{F}}{\partial p_1^2} = \frac{1}{\beta_c}\left[\frac{1}{p_1^2} + \frac{1}{p_1}\right]$
$\frac{\partial^2 \mathrm{F}}{\partial p_k^2} = \frac{1}{\beta_c}\left(\frac{1}{p_k}\right)$ for all $\tau_k \in \mathcal{T}$
$\frac{\partial^2 \mathrm{F}}{\partial z^2} = 0$

Then the Hessian is an $(M+1) \times (M+1)$ diagonal matrix with the entries given by $\frac{\partial^2 \mathrm{F}}{\partial p_1^2}, \frac{\partial^2 \mathrm{F}}{\partial p_k^2}, \frac{\partial^2 \mathrm{F}}{\partial z^2}$. Recall that β_c is positive since greater savings a_k (i.e., lower c_k) correspond to higher utility.

In addition, for more than one departure time alternative, all the multinomial choice probabilities $\{p_k : \tau_k \in \mathcal{T}\}$ are between $(0, 1)$.

Thus, all the diagonal entries of the Hessian are non-negative; this implies that the Hessian is positive semidefinite.

Since the Hessian is positive semidefinite, the objective is convex. □

The convex program 3.27–3.31 can be solved using open-source solvers such as CVXPY (Diamond and Boyd, 2016). After solving for the optimal probabilities $\{p_k : \tau_k \in \mathcal{T}\}$, determine the associated optimal cost $\{c_k : \tau_k \in \mathcal{T} \backslash \{\tau_1\}\}$ using Equation 3.36, where Equation 3.36 follows from Equation 3.33.

$$c_k^\star = c_1 - \frac{1}{\beta_c} \left[\log\left(p_k^\star\right) - \beta_d d_k - \log\left(p_1^\star\right)\right] \qquad \forall t_k \in \mathcal{T} \backslash \{t_1\} \tag{3.36}$$

3.5 Alternative Optimization Strategies

The formulation 3.27–3.32 takes the perspective of a service provider that aims to maximize revenue. For comparison, an alternative optimization strategy that focuses on system-level performance can be formulated. Observe that the objective in 3.27 has two components: an expression derived from revenue maximization (minimum savings), and wz which refers to reducing demand peaks. A revised optimization problem that drops the first revenue maximization component is shown in 3.37–3.42. This formulation minimizes the maximum increase in demand between one departure time and the next.

$$\min_{\substack{p_k:\tau_k\in\mathcal{T},\\ z}} z \tag{3.37}$$

$$\text{s.t.} \quad \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_{k+1}}\right] + \mathbb{E}\left[\delta_r^{\tau_{k+1}}\right]\right) - \left(\mathbb{E}\left[\mathrm{L}_r^{\tau_k}\right] + \mathbb{E}\left[\delta_r^{\tau_k}\right]\right) \leq z \qquad \forall \tau_k \in \mathcal{T}\backslash\{\tau_M\} \tag{3.38}$$

$$\sum_{\tau_k\in\mathcal{T}} p_k = 1 \tag{3.39}$$

$$0 \leq p_k \leq 1 \qquad \forall \tau \in \mathcal{T} \tag{3.40}$$

$$p_k \geq e^{\beta_d d_k} p_1 \qquad \forall \tau \in \mathcal{T}\backslash\{\tau_1\} \tag{3.41}$$

$$z \geq 0 \tag{3.42}$$

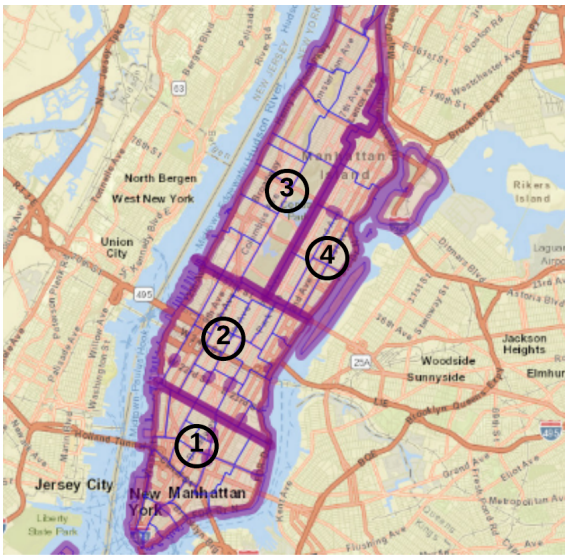

Figure 3.4: Manhattan divided into four regions.

3.6 Demonstrations & Network Analysis

In this section, we present experimental results using data from Lyft operations in Manhattan, NYC. We use data from rides that occurred on Friday December 14th, 2018

(NYCTLC, 2019) to estimate the model parameters. In addition, we limit the data to trips that started between 16:00-19:00 (local time) in the four regions that are shown in Figure 3.4. All rides in a zone are offered the same departure time alternatives and corresponding time-dependent surcharges. We use a rolling time horizon T of 50 minutes, and the users are offered five departure times that are evenly spaced out within the horizon (i.e., there is a difference of 10 minutes between successive departure time offers). Note that we consider the pricing intervals $[u_0, u_1)$ to be 10 minutes as well.

Our primary findings suggest that as the users value of time increases, the effectiveness of the peak-load pricing strategy decreases. In addition, to control lost revenue, the platform can adjust the weight parameter w. As w increases, the platform loses more revenue in favor of shaving peaks in the load process.

3.6.1 System model specification and rolling horizon implementation

At any pricing interval and associated future time horizon T, we use ride request received prior to $[u_0, u_1)$ to generate S_p^t and E_p^t. Then, we use the Manhattan ride request data to determine the maximum likelihood estimator of the upcoming arrival rates λ_{ij} between regions $i, j \in \mathcal{R}$. In addition, we use the ride duration of Manhattan trips to estimate an empirical service distribution $g_{ij}(\cdot)$ with CDF $G_{ij}(\cdot)$. The arrival rate and empirical service distribution are used to evaluate cumulative starts/ends $\mathrm{S}_f^t/\mathrm{E}_f^t$ associated with ride requests that will be received within the time horizon T.

In each region, after we evaluate the load process and determine the optimal prices that will be offered to users, we consider that the Manhattan ride requests that subsequently appear in $[u_0, u_1)$ to be ground truth observed data. Then, we probabilistically delay the start time of each observed ride based on the optimal MNL probabilities.

This process is successively repeated by first updating S_p^t and E_p^t to account for the choices of observed users $[u_0, u_1)$. Then, we analyze the subsequent pricing interval

$[u_1, u_2)$ and generate a new time horizon T that begins at u_2. Note that we also discount trips that start/end at u_1 from S_p^t/E_p^t since we are now only concerned with cumulative starts/ends in the new horizon $[u_2, u_2 + T]$.

3.6.2 Value of time and lost revenue

We analyze the impact of parameters β_c and β_d on the pricing strategy. For any specific departure time alternative τ_k, the change in utility is given as $\Delta V_k = \beta_c \Delta a_k + \beta_d \Delta d_k$. Setting ΔV_k to zero, we can evaluate the trade-off between savings and delay. In particular, $\Delta V_k = 0$ implies that $\Delta a_k = -(\beta_d/\beta_c)\,\Delta d_k$. Thus, in terms of the impact on utility, a unit increase in delay is equivalent to $-(\beta_d/\beta_c)$ in additional savings (recall that β_d is negative representing sensitivity to increased delay and β_c is positive representing sensitivity to greater savings). In other words, we can consider the value of time to be $\text{VOT} = -(\beta_d/\beta_c)$.

Maximizing the platforms revenue is equivalent to minimizing the expected user savings given by $\sum_{\tau_k \in \mathcal{T} \setminus \{\tau_1\}} a_k p_k$ (See Claim 3.4.2). The term $\sum_{\tau_k \in \mathcal{T} \setminus \{\tau_1\}} a_k p_k$ corresponds to the average lost revenue *per ride* based on the choices of the users. In Figure 3.5, we evaluate $\sum_{\tau_k \in \mathcal{T} \setminus \{\tau_1\}} a_k p_k$ for each region and then average the resulting sum across regions. We repeat the computation in Equation 3.43 at every pricing time interval and we plot the results for different VOT values.

$$\text{Lost Revenue} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \sum_{\tau_k \in \mathcal{T} \setminus \{\tau_1\}} a_k p_k \tag{3.43}$$

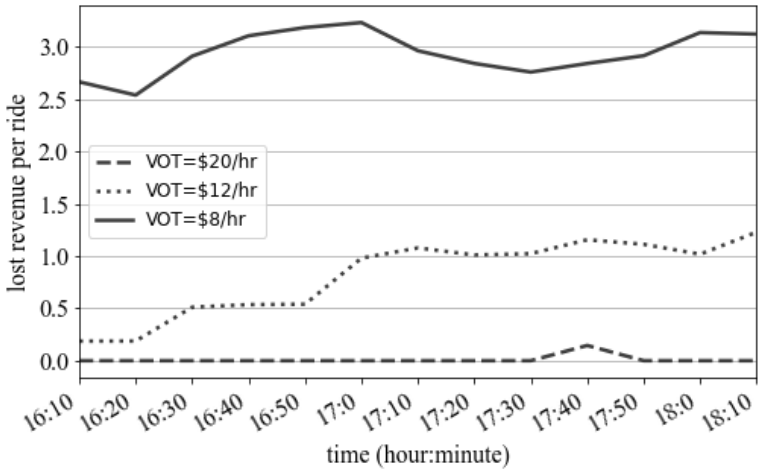

Figure 3.5: Lost revenue across time for different VOT values. The weight parameter w is set to one.

We observe that as the VOT decreases the average lost revenue increases. This indicates that users with a lower VOT are more likely to delay their departure time. In turn, delayed departure times result in losses to the platform. In contrast, users with high VOT choose to depart at earlier times and forgo the savings. To further incentivize high VOT users to delay the trip, the platform may increase the weight w to place greater emphasis on minimizing peaks in the load process as opposed to maximizing revenue.

In Figure 3.6, we illustrate the impact of the weight w on the lost revenue for a fixed VOT of \$12 per hour. We show that as the weight parameter increases in the optimization objective, the losses to the platform increase as well; this indicates that the platform prioritizes restricting peaks in load the process over generating revenue. On the other hand, when the weight is low, the lost revenue is negligible; this indicates that the platform does not offer users low cost departure time alternatives to avoid a decrease in its revenue.

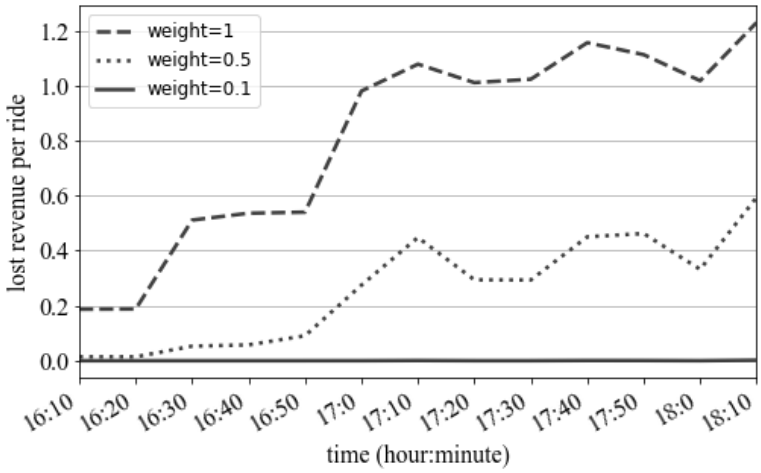

Figure 3.6: Lost revenue across time for different weight values. VOT is $12 per hour.

Figure 3.7 illustrates the difference in lost revenue when formulation 3.37–3.42 is used instead of 3.27–3.32. Formulation 3.27–3.32, which takes the platform's perspective (revenue+load), results in a lower level of lost revenue.

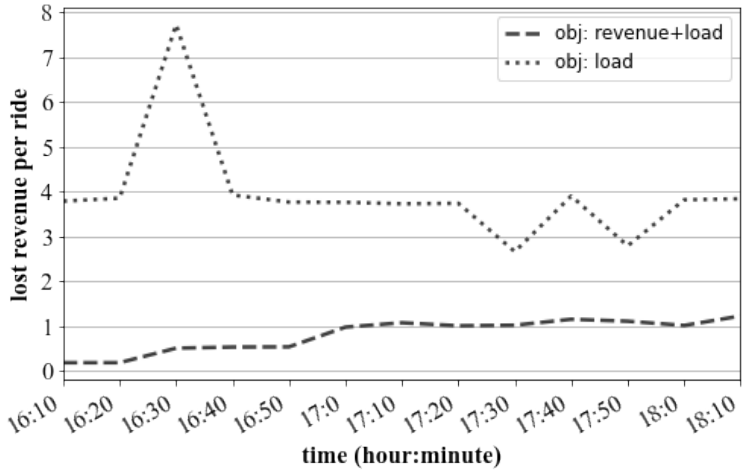

Figure 3.7: Lost revenue across time for different objectives: revenue maximizing vs. peak minimization only. VOT is $12 per hour. Weight parameter is 1.

3.7 Conclusion

In this chapter, we propose a pricing mechanism that limits peaks in demand to the available supply. In contrast to surge pricing, we offer user the option to delay their trip departure time in exchange for a reduced trip cost. Thus, by pricing different departure time alternatives, we aim to disperse users away from peaks in the load process, where an increase in the load process represents lost idle drivers.

As opposed to equilibrium-based methods that assume steady-state conditions, the proposed pricing mechanism focuses on the time-dependent system state and the associated transient probabilistic demand processes. In particular, we use a probabilistic characterization of future spatio-temporal demand to determine time periods with increased load. Then, we use the resulting load process to implement real-time pricing that reacts to the current and predicted system state.

In addition to restricting the load process, the pricing strategy aims to maximize platform revenue while representing user choices using a multinomial logit model. Simulation results using data from Lyft rides observed in Manhattan highlight the trade-off between maximizing revenue and restricting the load process. The results also exhibit the impact of user characteristics on the performance of the pricing strategy; specifically, we observe that as the users value of time increases, the effectiveness of the pricing strategy in terms of restricting the load process is diminished.

Chapter 4

E-Scooters in Austin, TX: Effect of Transit Network Redesign on E-Scooter Ridership

4.1 Introduction

In this chapter, we investigate the interaction between e-scooter and transit service in Austin, TX. In 2018, CapMetro, the local transit agency, implemented a major redesign of their transit network. This redesign–CapMetro Remap (CapRemap)–involved restructuring the transit service towards a high frequency network. Shortly before CapRemap, e-scooters were introduced in Austin and their ridership experienced a steady growth as shown in Figure 4.1. The objective of this research is to study the change in e-scooter ridership that can be attributed to CapRemap. This analysis would help understand the trade-offs between transit and e-scooters.

The primary difficulty in isolating the impact of CapRemap on e-scooter ridership results from the existence of confounding variables. In more detail, observing increased ridership in certain areas may be attributed to demographic variables or to the proximity to the UT Austin campus. The statistical analysis proposed in this section controls for such confounding variables using a matching approach. We isolate areas that were impacted by CapRemap and match them to reference areas that were not impacted. Then, we analyze the *trend* in scooter ridership across the matched areas. In other words, we compare the *change* in ridership of the control group to the corresponding change observed in the reference group.

The matching relies on sociodemographic variables and a proximity to UT measure. The demographic variables used are population density, retail employment, me-

dian income, and proportion of young adults (under 34). The assumption is that areas that match on those variables exhibit a common trend regarding the change in ridership. Thus, if the control area shows significant trend deviation relative to the reference area, this could be a result of the intervention (CapRemap).

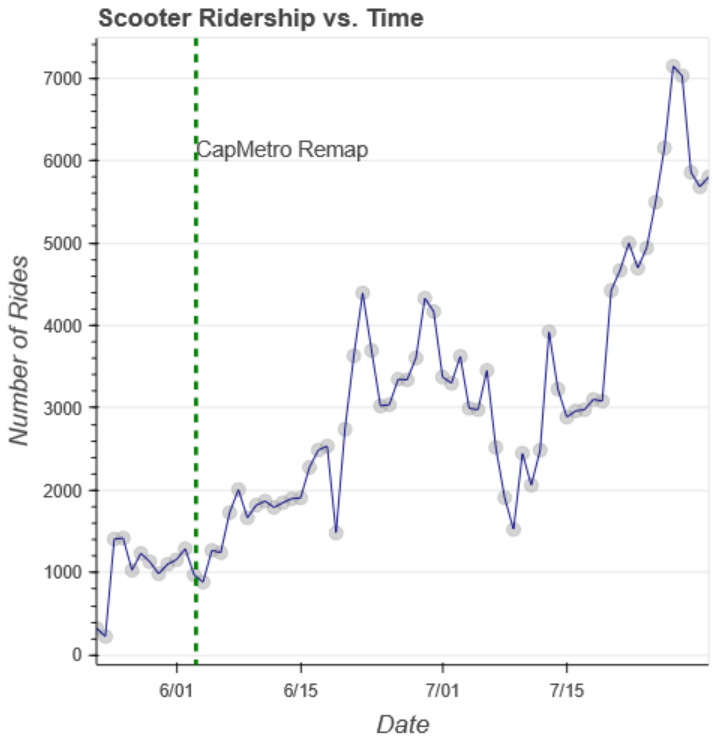

Figure 4.1: E-Scooter ridership across time in Austin, TX.

4.1.1 CapRemap change in bus service

To visualize the impact of CapRemap, Figure 4.2 illustrates the change in bus service throughout Austin. As observed, several areas lost frequent service, including ones that have a high proportion of minorities that would rely on transit as a primary mode of transport. However, it is not immediately apparent whether e-scooter trips can replace transit in areas that lost service or complement transit in areas that gained high frequency lines. In fact, figure 4.3 shows that scooter ridership is heavily concentrated in specific

areas close to downtown or the UT Austin campus.

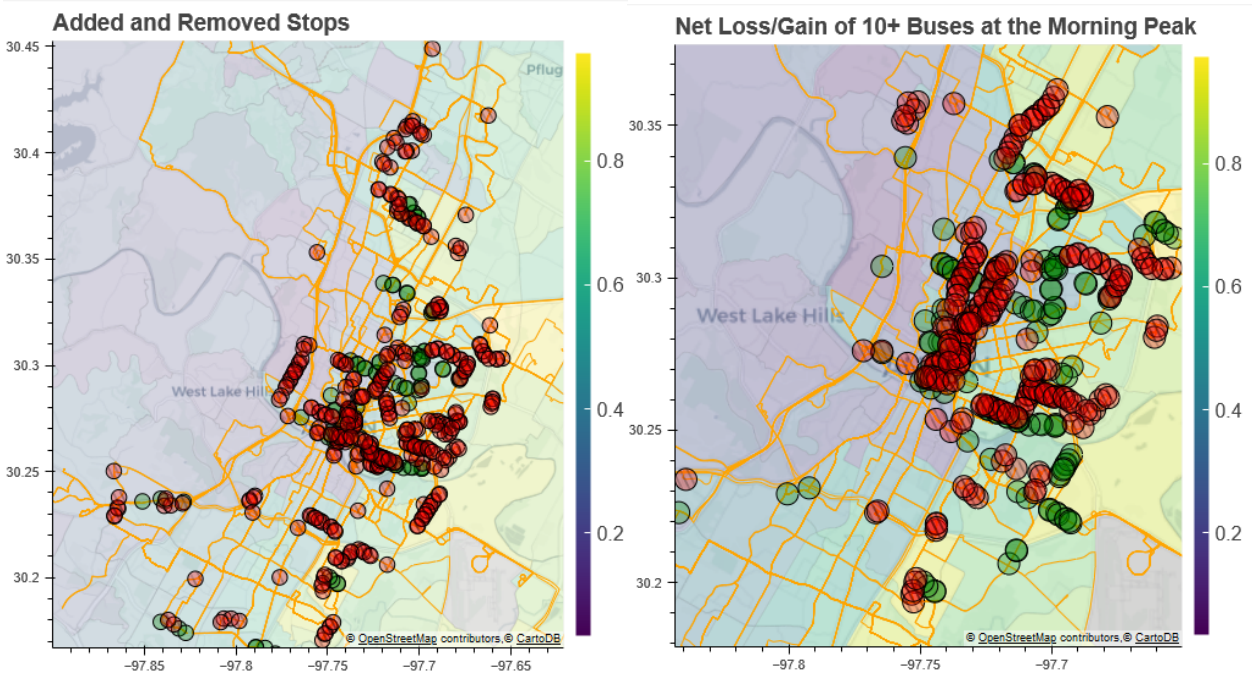

Figure 4.2: Added (green) and removed (red) stops following CapRemap, and stops that had a net loss or gain of more than 10 buses during the morning peak. The color bar shows the proportion of minorities in each census tract.

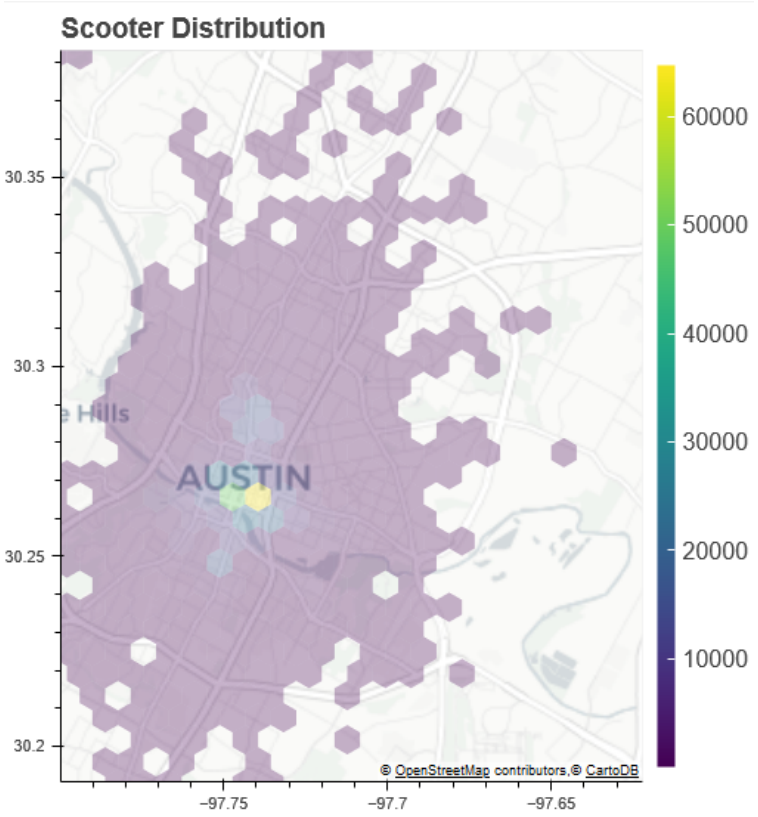

Figure 4.3: The geographic distribution of e-scooter rides in Austin, TX. The color bar represents number of rides.

4.2 Identification of Traffic Analysis Zones Impacted by CapRemap

Before we proceed to matching and estimating the effect of CapRemap on scooter ridership, we first start by identifying the impact area. In what follows, the geographic analysis unit is taken to be a traffic analysis zone (TAZ), where TAZs are defined by CAMPO (the planning agency in the Austin area).

First, we need to map the transit service change from bus stop level to TAZ level. Figure 4.4 shows one approach for this mapping based on the proportion of each bus stop *buffer* in the TAZ. The bus stop buffer has a 1/4 mile radius around the stop, where this 1/4 mile distance represents the 85th percentile walking distance to stops. In other words,

the 1/4 radius buffer is an appropriate representation of the stop coverage area. In the illustrated example, a bus stop with a net gain of 31 stops has 25% of its 1/4 mile buffer area in the TAZ and another with a net loss of 10 bus stops has 50% of its buffer area in the TAZ, the resulting TAZ score is an area-weighted impact of each bus stop.

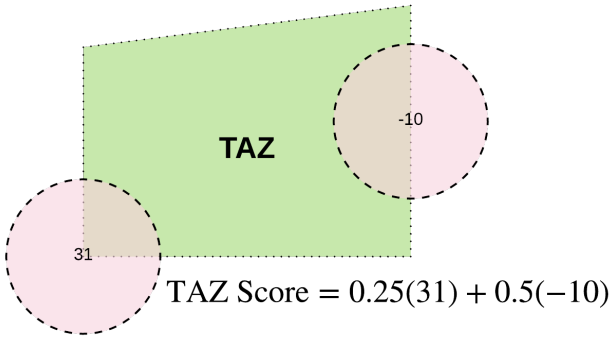

Figure 4.4: Mapping bus stop service change to TAZ level service change.

Figure 4.5 shows the TAZ level score across the network. As shown, higher values indicate improved service while negative values indicate adversely impacted TAZs. We use that TAZ level score to isolate areas that were impacted by CapRemap. To define the areas that were significantly impacted by CapRemap, we look at the histogram of TAZ scores. We observe that a threshold of 30 is a reasonably extreme value such that TAZs with a score greater than 30 had significant improvement and those with a score less than -30 had a significant reduction in service.

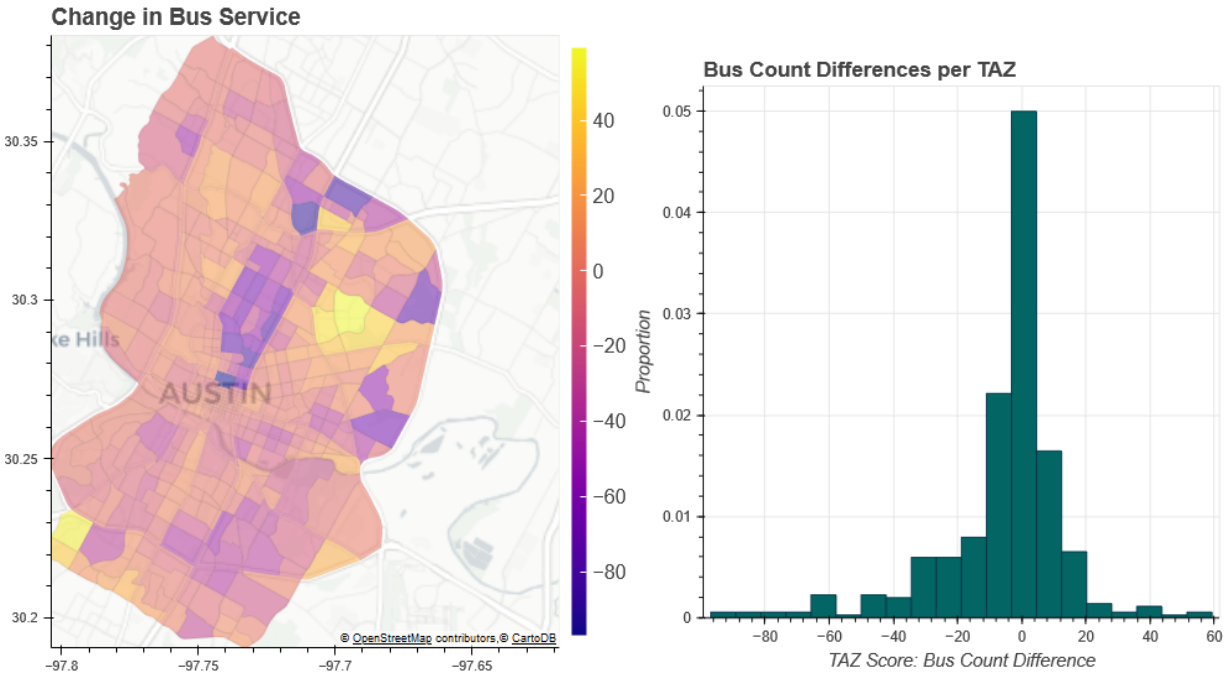

Figure 4.5: TAZ score illustrating bus service change impact at the TAZ level. Histogram of service changes across TAZs.

Subsequently, after identifying TAZ scores that are more extreme than the threshold, we can *group* a set of impacted TAZs as shown in Figure 4.6. A label of 'Negative' indicates that the area had a reduction in service and a label of 'Positive' indicates that the area had improved service. These significantly impacted areas are roughly around Central, East, and South Austin.

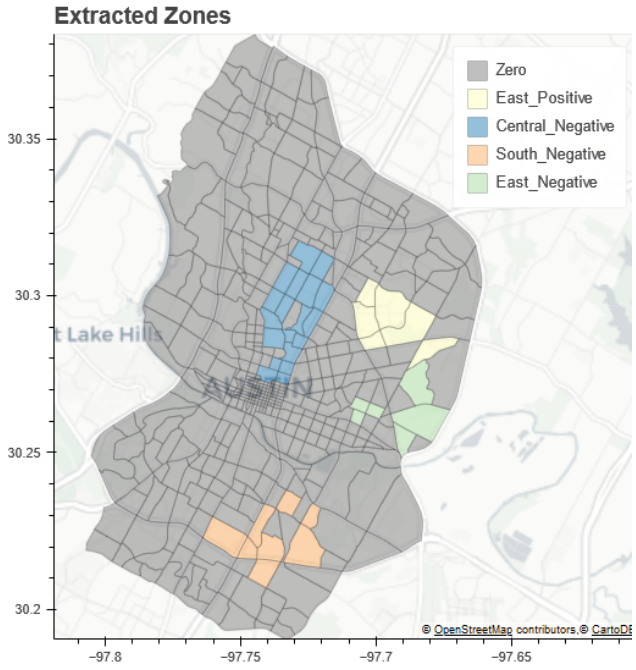

Figure 4.6: Areas that are either adversely or positively impact by CapRemap. Those areas represent a group of TAZs that had significant changes in bus service.

4.3 Matching Impacted TAZs to Reference Zones

As previously mentioned, to match the impacted TAZs to reference zones we use a set of demographic variables and the proximity of a TAZ to the UT main campus. The demographic variables vary significantly across Austin as shown in Figure 4.7.

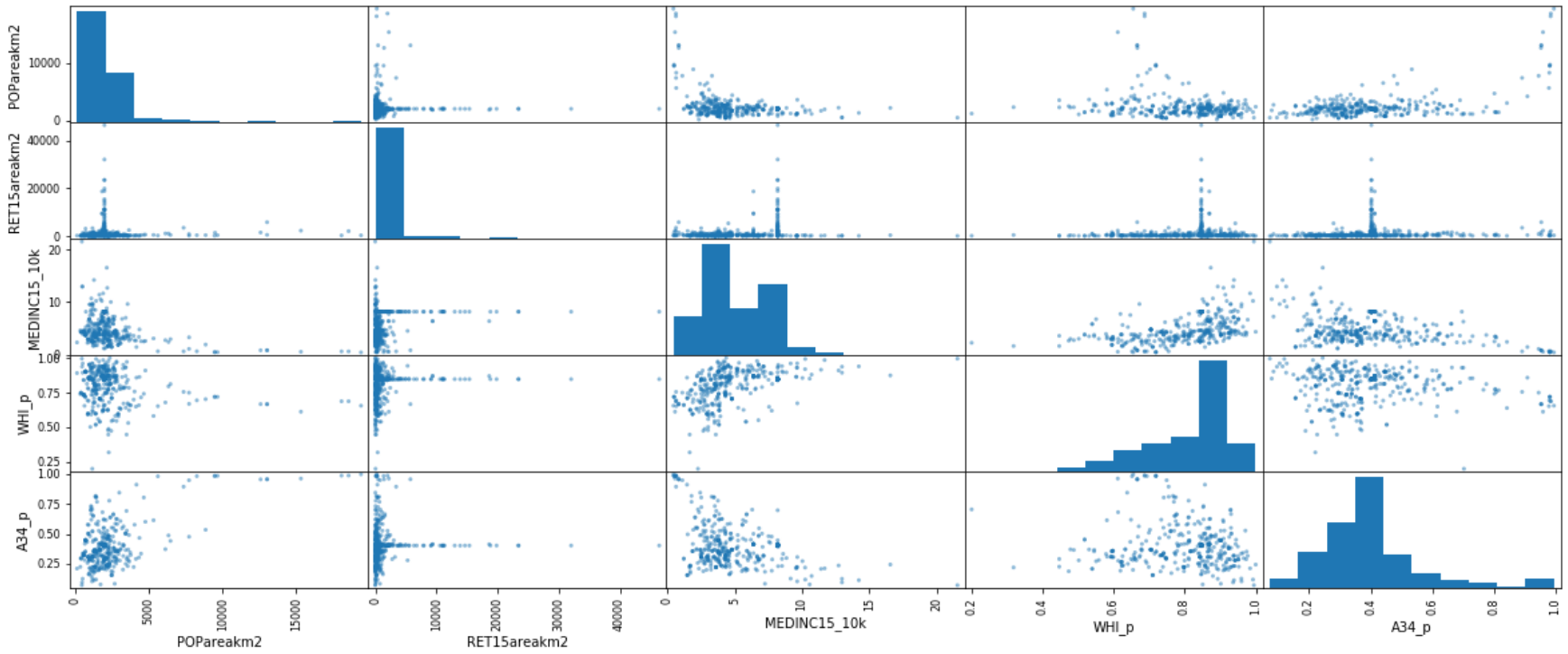

Figure 4.7: Distribution of demographic variables across traffic analysis zones in Austin. A34_p: 2016 proportion of young people under the age of 34. WHI_p: 2016 proportion of White people. MEDINC15_10k: 2015 median income in units of $10,000. RET15areakm2: 2015 retail employment per unit area. POPareakm2: 2015 population density per unit area.

4.3.1 Mahalanobis distance matching

Given the relevant demographic variables, the Mahalanobis distance measure is used to find the similarity between TAZs. The Mahalanobis distance is one approach for matching and it is defined in Equation 4.1. Rubin (2006) provides a detailed discussion of different matching procedures used for causal analysis in observational studies.

$$D = \sqrt{(u - v)V^{-1}(u - v)^T} \tag{4.1}$$

The term u is a vector of demographic variables and proximity to UT for a TAZ, the term v is the corresponding vector for a reference TAZ. The matrix V^{-1} is the inverse of the covariance matrix of the features, where this matrix is used to normalize for the scale of different variables. Thus, the Mahalanobis distance measures how close TAZs are to each other in terms of sociodemographic and UT proximity variables while normalizing for the scale of each variable.

The Mahalanobis distance can be used to match one impacted TAZ to a reference TAZ. Since each impacted area is composed of multiple grouped TAZs (Figure 4.6), the proposed matching approach proceeds to find a reference TAZ for each TAZ within the impacted area. The aggregation of the matched TAZs forms the reference *matched area*. Figure 4.8 shows the matched area (collection of green TAZs) for the negatively impacted central Austin area shown in red. Similar matched reference areas can be found for other impacted locations (Figure 4.6).

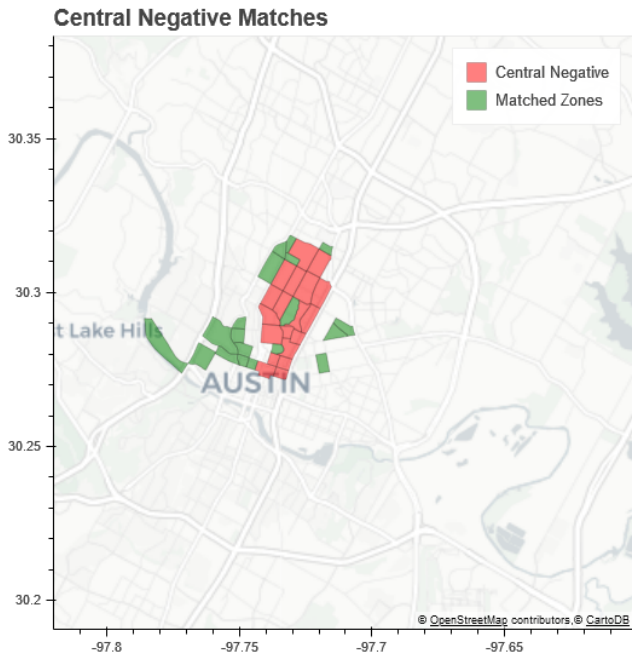

Figure 4.8: In red we have an area adversely impacted by CapRemap. In green is a matched reference area with similar demographics but not affected by CapRemap.

4.4 Difference in Differences Statistical Analysis

After matching, we proceed to implement the difference-in-differences regression to infer the impact of CapRemap on scooter ridership. The difference-in-differences (DID) procedure is commonly used in natural experiments (quasi-experiments). In particular, DID models can be used for estimating the effect of an intervention where before-after data is available (Hill et al., 2018).

Validity of the difference-in-differences approach is subject to the standard causal analysis assumptions in non-random (natural) experiments. Of those assumptions, for the specific case of this CapRemap analysis, the stable unit treatment value assumption (SUTVA) may not hold across matched pairs. SUTVA implies that the impact of the CapRemap intervention should not spill over to control group TAZs. This assumption

may not be necessarily true since the network redesign can impact non-adjacent TAZs as a consequence of origin-destination trip patterns.

Another assumption that is specific to DID models is referred to as the parallel trends requirement. This assumption states that, in absence of the intervention (CapRemap), the change in ridership at impacted areas would have paralleled the trend in reference areas. In other words, the observed change in the control group is a proper counterfactual for the change in the impacted group, where counterfactual refers to the scooter ridership in the impacted area had the intervention not happened.

The parallel trends assumption is best illustrated in Figure 4.9. In this figure, the control group outcome y starts at A and ends up at E after the intervention. The impacted group starts at B and ends up at C. The parallel trends assumption states that, had the intervention not occurred, the change in outcome for impacted group would have paralleled the control group change. The counterfactual for the impacted group, or imagined outcome had the intervention not happened, would be point D. This implies that difference δ between C and D is the intervention effect.

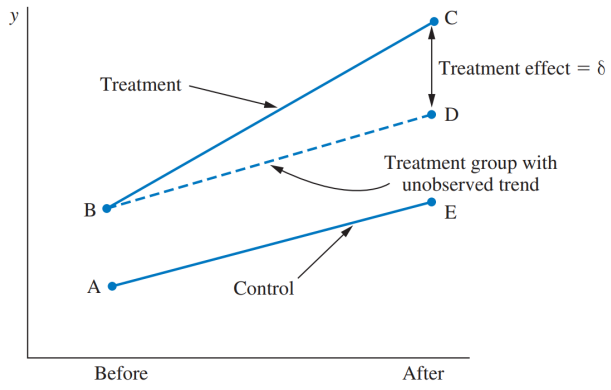

Figure 4.9: The difference-in-difference parallel trends assumption. Extracted from Hill et al. (2018).

The intervention effect is estimated using a regression model. For the scooter rider-

ship application, the regression takes the form shown in Equation 4.2. In this regression, 'date' is a boolean that is 1 post-intervention and zero otherwise, 'change' is a boolean that is 1 for impacted areas and zero for control. The dependent variable 'count' refers to the number of scooter rides in one day.

$$\text{count} = \beta_0 + \beta_1(\text{date}) + \beta_2(\text{change}) + \beta_3(\text{change}) * (\text{date}) \tag{4.2}$$

The parameters of the model are illustrated in Figure 4.10. The parameter β_0 reflects the expected number of rides in the control area and before CapRemap. The parameter β_1 refers to the trend; to be precise, β_1 captures the change in scooter ridership that can be attributed to time alone (i.e., does not include any additional effect of the CapRemap intervention). The parameter β_2 refers to the change in ridership that is attributed to being in the impact area relative to the control zone; similarly, β_2 does not factor in any additional ridership differences that result from the CapRemap intervention. The parameter β_3 refers to the difference-in-differences intervention effect.

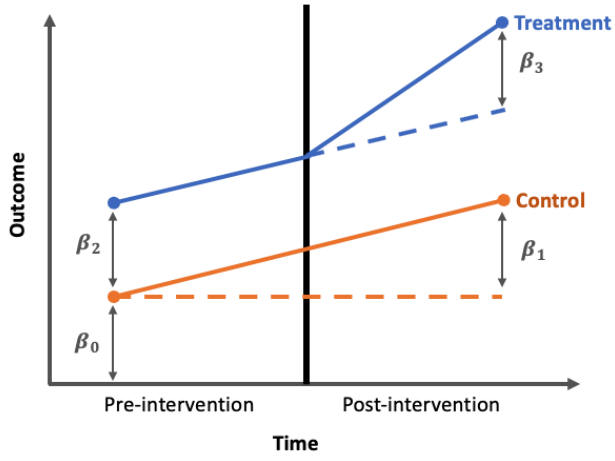

Figure 4.10: The difference-in-difference regression.

4.4.1 Data preparation

To estimate the model, we create four different data-sets for every pair of matched zones. A sample of one dataset is shown in Table 4.1. For each day (a data point), we store the scooter use as the variable 'count', and we have two boolean variables 'change' and 'date'. The variable 'change' is one if the area is impacted by CapRemap and zero for reference zones. The variable 'date' is one if the data point is take after CapRemap and zero otherwise.

Note that we only consider data collected before June 19th, 2018. CapRemap was implemented on June 3rd. The scooter data available starts at May 23rd, 2018. Thus the data consist of about 11 days before CapRemap and 16 days after its implementation. We do not use data beyond June 19th to avoid exogenous factors that may have emerged during that time frame.

Table 4.1: Sample data for a matched pair of zones

day	count	change	date
2018-05-23	32	1	0
2018-05-24	13	1	0
2018-05-25	115	1	0
2018-05-26	120	1	0

4.4.2 DID results

After fitting the difference-in-difference regression for each matched pair, we obtain an ordinary least squares estimate for the parameters. The results for the central Austin region (Figure 4.8) are shown below in Figure 4.11. Observe that the coefficient β_3 had a value of 67. Moreover, for the null hypothesis that $\beta_3 = 0$, we get a t-score of 2.75 which implies that we reject the null hypothesis under α levels as low as 0.008. The 95%

confidence interval for the coefficient value is $(18.08, 115.99)$. Overall, the regression has a reasonably high adjusted R-squared of 0.637.

OLS Regression Results

Dep. Variable:	count	**R-squared:**	0.637
Model:	OLS	**Adj. R-squared:**	0.615
Method:	Least Squares	**F-statistic:**	29.24
Date:	Tue, 12 Oct 2021	**Prob (F-statistic):**	4.60e-11
Time:	21:50:47	**Log-Likelihood:**	-278.90
No. Observations:	54	**AIC:**	565.8
Df Residuals:	50	**BIC:**	573.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	38.1818	13.268	2.878	0.006	11.532	64.832
change	45.2727	18.764	2.413	0.020	7.584	82.961
date_bool	33.0057	17.236	1.915	0.061	-1.613	67.625
change:date_bool	67.0398	24.375	2.750	0.008	18.081	115.998

Omnibus:	6.713	**Durbin-Watson:**	1.637
Prob(Omnibus):	0.035	**Jarque-Bera (JB):**	8.041
Skew:	0.397	**Prob(JB):**	0.0179
Kurtosis:	4.716	**Cond. No.**	7.58

Figure 4.11: The difference-in-difference regression results for the central Austin negatively impacted area.

Thus, for the central Austin region shown in Figure 4.8, the results indicate that there is a significant change in scooter use after CapRemap, where this change in scooter use is not explained by other sociodemographic variables or features. This analysis was repeated for other zones; however, for none of the other zones were we able to reject the null hypothesis that $\beta_3 = 0$. This indicates that the results are inconclusive and that the DID parameter is in fact *not* significant in most cases.

Recall that the model has limitations that may result in the inconclusive results. In

addition to violation of the SUTVA assumption discussed earlier, the model can not decouple the effects of using scooters as a first-mile last-mile service from the use of scooters in competition with transit. Thus, the coefficients may absorb different effects simultaneously which leads to misleading results.

4.5 A Note on Equity in Transit Planning

While e-scooters did not sufficiently replace transit service, it is not clear if minority groups were disproportionately impacted by CapRemap. In fact, CapRemap raised several equity concerns and accusations of racial discrimination. Despite CapMetro's service equity analysis showing compliance with FTA's policies, activists are still determined that the redesign violates Title VI requirements. This section briefly discusses the impact of CapRemap across sociodemographic groups in Austin, TX. In particular, we focus on limitations of commonly used equity analysis procedures that comply with the FTA's Title VI requirements.

4.5.1 Limitations of current FTA-compliant equity analysis methods

Before analyzing the data in greater detail, it is worth mentioning CapMetro's equity analysis that showed compliance with Title VI. In fact, CapMetro states that the benefit to minorities from the service adjustments far exceeds the potentially adverse impact.

As discussed in an MPO policy meeting, CapMetro evaluated each of the major service adjustments by studying the demographics of a 1/2 mile walk-shed that surrounds the changed routes. The key analysis approach is to first find out whether the % minority population around the walk-shed is greater than the average % minority population in the total service area. If that is the case, proceed to determine whether alternative routes can cover the minority block groups that lost service. This route level analysis is

consistent with the FTA's Title VI requirements. CapMetro's results shows that most areas with lost service will be covered by alternative routes and in many instances there will be new high-frequency options as well.

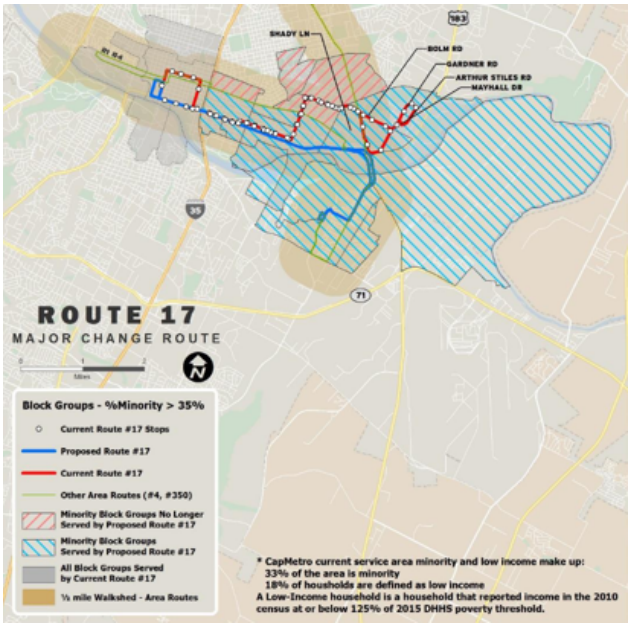

Figure 4.12: Sample from CapMetro's route-based analysis. Source: CAMPO transportation policy meeting.

The limitations of CapMetro's route-level equity analysis are as follows:

1. While it is often mentioned that there will be high frequency routes close to minority groups (similar to other studies that focused exclusively on those high frequency routes), a detailed service frequency analysis seems to be lacking. The addition of high frequency routes does not give the full picture of service changes on frequent and non-frequent routes

2. Forming 1/2 mile walk-sheds around routes (with 1/4 mile strips on each side) is a common method for measuring system coverage. Despite that, passengers board their buses at stops, and the 1/4 mile distance is based on the 85th percentile walk-

ing distance to those stops. Wouldn't stop-based coverage, with a 1/4 mile radius around stops, be more appropriate in that case?

3. The routes analyzed were restricted to those that had a greater than 25% change in geographic coverage or service characteristics, where this 25% threshold was set by CapMetro. Even after selecting the routes with major changes, they were only analyzed further if the % minority population in the walk-shed was greater than 35%. Does this exclude parts of the network that were adversely impacted?

4. As shown in the Figure 4.12, the stops at Gardner Rd and Arthur Stiles Rd are removed, and their location will no longer be within a 1/2 mile walk-shed of any route. However, the minority block group in which they are located (blue) is assumed to be covered by the adjusted route. Clearly, passengers that previously used those stops will no longer be a short walk away from any transit line; but, they are considered to be covered due to the irregular shapes of census block groups. In particular, the analysis assumes that a block group has transit service if any part of its area overlaps with the walk-shed. A better equity analysis approach would restrict coverage to the area within a 1/4 mile distance from bus stops.

4.5.2 A peak-hour stop-based analysis approach

Focusing on the weekday morning peak service (7–10 a.m.), which targets essential home-based work trips, we implement a *stop-level equity analysis* of the service changes.

The proposed stop-based approach in this section contrasts with the previous CapMetro analysis as follows: (1) The change in frequency is evaluated at each stop by measuring the difference in doors opening before and after CapRemap (2) A buffer with a 1/4 mile radius is created around each stop to determine the demographic characteristics of affected riders (3) The approach includes changes to all routes — not just ones that pass

CapMetro's thresholds for significant changes and disparate impact (4) The impact of the service change is restricted to the population within the buffer to avoid irregularities in census data and to accurately represent the coverage area.

Figure 4.2 shows the stops that were added or removed, where the color bar shows the proportion of minorities in each census tracts. It is evident that many stops were removed in areas with a high proportion of minorities. That said, looking only at new or removed stops is not representative of the full service change. The removal of a stop that had low service is highlighted while major service reductions at other stops are not shown. Similarly, in their transition to a high frequency network, CapMetro may have significantly improved the frequency at existing stops without adding many new stops. The second map in Figure 4.2 shows stops that experienced a change of more than 10 buses during the morning peak; this better illustrates areas with significant changes.

That said, for a precise analysis of the change in service frequency and its impact on different demographic groups, some detailed stop-level demographic information is needed. We proceed by computing the demographic characteristics *per stop*. Then, using data on changes to bus frequency at each stop, we define aggregate metrics that describe the change in level of service experienced by each demographic group.

Stop-level demographic data

Getting stop-level sociodemographic data requires projecting variables from census tracts to the stop buffers. To do so, we can use the proportion of the buffer that lies in each tract. This mapping is best illustrated in an example. Figure 4.13 shows how demographic variables are computed for a buffer that overlaps with two tracts (25% of the buffer area is in tract 1). The term 'inter. area' refers to the area of intersection between the buffer and the tract.

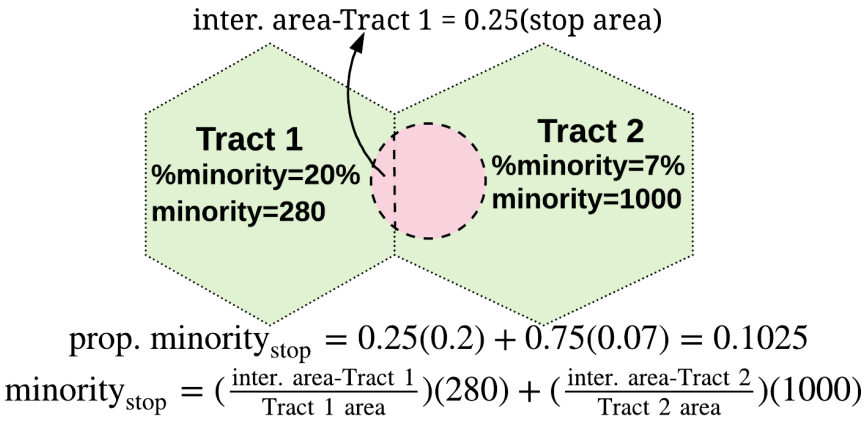

Figure 4.13: Mapping census tract demographic data to stop level data.

In general, the number of minorities and the proportion of minorities within the 1/4 mile buffer can be determined as in Equations 4.3 and 4.4. Similar equations can be used to map any census demographic data to stop-level data.

$$\text{prop. minority}_{\text{stop}} = \sum_{\text{tracts}} \left(\frac{\text{inter. area}}{\text{stop area}} \right) (\text{prop. minority})_{\text{tract}} \tag{4.3}$$

$$\text{minority}_{\text{stop}} = \sum_{\text{tracts}} \left(\frac{\text{inter. area}}{\text{tract area}} \right) (\text{minority})_{\text{tract}} \tag{4.4}$$

Stop-level service change metrics

After computing the stop demographic data, we define stop-level service metrics. Specifically, let's define 'impact' to be the change in service after implementation of CapRemap. From that, 'doors opening' is defined as the impact at stops with improved service, and 'doors closing' is defined as the impact at stops with reduced service.

$$\text{impact} = (\text{no. buses post-CapRemap}) - (\text{no. buses pre-CapRemap}) \tag{4.5}$$

$$\text{doors closing} = \max\{-\text{impact}, 0\} \tag{4.6}$$

$$\text{doors opening} = \max\{\text{impact}, 0\} \tag{4.7}$$

Aggregate impact metrics

Given the stop-level demographic data and metrics, we can now define aggregate metrics that accurately describe the service changes per demographic group.

$$\text{Expected Impact} = \frac{\sum_{i \in \text{stops}} (\text{impact})_i (\text{minority})_i}{\text{total minority in coverage area}} \tag{4.8}$$

$$\text{Frac. DO} = \frac{\sum_{i \in \text{stops}} (\text{doors opening})_i (\text{prop. minority})_i}{\sum_{j \in \text{stops}} (\text{doors opening})_j} \tag{4.9}$$

$$\text{Frac. DC} = \frac{\sum_{i \in \text{stops}} (\text{doors closing})_i (\text{prop. minority})_i}{\sum_{j \in \text{stops}} (\text{doors closing})_j} \tag{4.10}$$

The 'Expected Impact' is the average service change experienced by a minority person. In other words, if a minority person was sampled at random from the service area, this is the change in service that they will experience.

The 'Frac. DO' is the fraction of service improvements that went to minorities. Similarly, 'Frac. DC' is the fraction of service reductions inflicted on minorities. In contrast to the 'Expected Impact' metric, those measures are not dependent on the density of minorities in a particular area. For example, greatly improving service in a location that is dense with minorities while leaving out many minority areas unconnected would give a large positive 'Expected Impact', but this may be undesirable.

4.5.3 Stop-level equity results & discussion

Table 4.2: Stop-level aggregate metrics for CapRemap

	Expected Impact	**Frac. DO**	**Frac. DC**
Minority	-5.54	0.55	0.52
White	-6.10	0.45	0.47
Black	-4.25	0.098	0.078

The results in Table 4.2 show that, on average, Austin's residents would see fewer buses passing during the morning peak! While CapRemap added frequent lines, this was at the expense of other non-frequent service. If we sample a minority person at random, we would find that she experienced a net loss of around 5 buses passing during the morning peak.

However, in terms of equity, there does not seem to be any bias against minorities. The fraction of service improvements that went to areas with Black people was low (only 9.8% of the total service improvements). At the same time, at 7.8%, the fraction of service reductions that was inflicted on areas with Black people was also low. Overall, minority areas were allotted 55% of the total service improvements (doors opening) and they received 52% of the total service reductions. Meanwhile, areas with White people were allotted 45% of the total service improvements and they received 47% of the total service reductions.

The results indicate that minority areas did not simultaneously receive a lower fraction of the service improvements and a greater fraction of the service reductions, which indicates that there is no apparent bias against minorities in the distribution of service modifications.

In response to complaints by activists, the FTA stated that the total minority population close to frequent service substantially increased. They cited the fact that 50,000 additional minority persons will be close to such frequent service. However, the focus on the increase in frequent service may be misleading since people observed fewer buses on average.

4.6 Conclusion

This section looked at the impact of CapRemap, Austin's transit network redesign, on scooter ridership in the city. To isolate the impact of CapRemap on scooter use, we implement a difference-in-differences (DID) statistical analysis. First, we identify the TAZ's that experienced significant changes in bus service due to CapRemap. Then, we group blocks of such TAZs into impacted areas and we *match* the resulting impacted areas to reference/control areas. The matching procedure is used to control for confounding variables, and it finds reference areas that have similar demographic characteristics to the impacted areas. In particular, the matching approach uses the Mahalanobis distance to find pairs of similar areas between reference and control groups. After matching is complete, the DID model is used to estimate the difference-in-differences effect. The DID model assumes that the matched areas would have parallel scooter ridership trends had CapRemap not occurred. Thus, the difference-in-differences CapRemap intervention effect is estimated by quantifying the deviation from the parallel trends assumption within the impacted group. The results are inconclusive. While we found that CapRemap increased scooter use near the UT campus where transit service was reduced, this result did not hold in other parts of Austin. In addition, there may be other factors that led to violation of causality assumptions and biased the results.

Section 4.5 of this chapter discusses the equity analysis used by CapMetro to satisfy the FTA Title VI requirements. Limitations of CapRemap's equity analysis are discussed, and an alternative stop-based procedure that addresses existing deficiencies is proposed. Although CapRemap led to accusations of racial discrimination, we do not find any bias against minorities in the distribution of service changes. However, while CapMetro claimed increased benefit to minorities and emphasized the added high-frequency routes, we found that minority groups were overall worse off after the network redesign.

Chapter 5

Conclusion

This dissertation explores the management and operation of on-demand mobility systems. Chapters 2 and 3 analyze inefficiencies in the operation of ridesourcing services and propose strategies for supply/demand management. Chapter 4 investigates e-scooter service in Austin TX and the impact of transit network redesign on scooter ridership.

5.1 Contributions

5.1.1 Ridesourcing

For ridesourcing systems, the contributions are as follows:

1. In contrast to equilibrium-based and steady-state stochastic methods, we use time-varying models derived from transient analysis of queueing systems.

2. Based on those time-varying models, we analyze control policies aimed at managing driver supply and maintaining a desired reach-time level of service. In the context of reservations, where an admission control policy prioritizes book-ahead rides, we provide a time-dependent upper bound on the probability of reach time violation for non-reserved rides.

3. Given this probability of reach-time violation under the control policy, we determine the target number of drivers that needs to be provided in each region. Effectively, this target limits the probability of reach-time violation to be within a desired tolerance.

4. We use the targets in a min-cost flow reformulation of the driver dispatching and rebalancing problem. The optimal solution for this program represents the minimum number of *idle* driver transitions between adjacent regions that is needed to maintain the targets.

5. To manage demand, we propose a peak-load pricing strategy that gives users incentives to delay their trips during periods of high demand. Similar to the reservations model, we use a time-dependent characterization of the system dynamics. However, unlike the reservations study where the time-varying probability of reach-time violation was evaluated, the stochastic processes are only analyzed in expectation. That said, the processes representing expected number of trip starts and ends are still time-varying functions.

6. In the peak-load pricing analysis, we use multinomial logit (MNL) models to represent the user choice among departure time alternatives. For each alternative, the MNL model gives the probability of users choosing the alternative given the associated trip cost and delay. The optimization program developed takes the platforms perspective, and it aims to maximize revenue subject to demand shaving constraints and the MNL user choice probabilities. A key component of this dissertation is showing that the resulting non-convex optimization program reduces to a *convex* equivalent. This convex reformulation relies on expressing the optimization program in terms of choice probabilities instead of trip cost.

5.1.2 E-Scooters

For the e-scooter ridership and transit network redesign study, the contributions are as follows:

1. We propose a procedure for evaluating the impact of transit network redesign on e-scooter ridership. Using data from CapRemap, Austin's network redesign in 2018, we estimate the change in scooter ridership that can be attributed to CapRemap. The primary challenge is isolating the effect of CapRemap from other confounding variables that may influence ridership. To do so, we implement a Mahalanobis distance matching approach that pairs areas impacted by CapRemap with reference/control areas. This matching uses demographic variables and proximity to the UT campus as indicators of similarity between different locations. Then, for every matched pair of areas, we use a difference-in-differences (DID) regression to estimate the effect of CapRemap. The DID model depends on the assumption of parallel trends; this assumption states that, in the absence of CapRemap, the change of scooter ridership in impacted areas would have paralleled that in reference areas. Then, the CapRemap effect is inferred as any significant deviation from this trend. Since the matched pairs are similar on demographic variables and the DID model relies on trends, the estimated CapRemap effect is free from time-related or demographic factors that may otherwise bias the results.

2. We investigate existing approaches used to meet the FTA's Title VI requirements regarding equity in transit network redesign. We highlight limitations of existing procedures, and we propose an alternative stop-based equity analysis. The stop-based analysis projects census-level demographic variables to the stop-level, and then uses aggregate metrics that better capture the impact of service changes on each demographic group.

5.2 Results

For reservations in ridesourcing systems, our results indicate that increased reservations lead to lower targets and subsequently a fewer number of idling drivers. In other words, the information provided by reservations helps us provide the appropriate number of drivers to fulfill book-ahead rides and limit the probability of reach-time violation for non-reserved rides. On the other hand, in the absence of reservations, to guarantee the same level of service, an increased number of drivers must be deployed and this leads to inefficiencies (excess idle drivers).

In terms of peak-load pricing, our results indicate that the user's value of time (VOT) has a large effect on the success of the demand shaving strategy. If the user's value of time is high, they are less willing to delay their departure time in exchange for a reduced fare. When the VOT is low, more users delay their trip by choosing lower cost trips and resulting in better demand shaving at the expense of some lost revenue.

The e-scooters study shows that, in many cases, the impact of transit network redesign on e-scooters ridership is insignificant. The results suggest that e-scooters did not replace transit in areas that lost service. Regarding equity implications of the CapRemap transit network redesign, our proposed stop-level analysis shows that there is no apparent bias against minorities. That said, on average, if we sample a minority person at random, we would find that they experienced a net loss in service. This result contradicts CapMetro's emphasis on benefits to minority populations through increased proximity to high frequency transit lines.

5.3 Future Work

While the majority of existing studies focus on equilibrium/steady-state analysis of emerging mobility services, existing research suggests that temporal variations in

supply/demand occur rapidly, and systems describing mobility services may not attain a steady-state equilibrium (Braverman et al., 2019). This rapid temporal variation in parameters suggests that there is a need for additional research on time-varying models and state-dependent control policies. Equilibrium-based methods are useful for high-level long-term planning; however, the nature of rapid and dynamic mobility systems requires innovative management strategies that react to the time-varying state of the system. The stochastic nature of supply/demand also warrants additional research that further investigates the impact of uncertainty on operational inefficiency and the analysis of control policies in light of this stochasticity.

In the transportation literature, equity remains one area that requires additional research. Across the transportation planning stages, from survey design to mode choice and network analysis, little consideration is given to minority groups and their behavior. It is well known that minority groups respond to surveys differently, use particular modes, and are often adversely impacted by large transportation projects. Additional research is needed on (1) procedures to incorporate their opinions and behavior into transportation planning, and (2) experimental and data-driven metrics that quantify the impact of transportation projects on travel choices of minorities. In this dissertation, we show how transit network redesign is one area that would benefit from further equity research.

Another critical research area is the safety and security of emerging mobility systems (Perrine et al., 2019). Apart from connected and autonomous vehicles, significant advances are being made in the monitoring of traffic and transportation infrastructure (Yahia et al., 2021a). This abundance of real-time data is essential for improving the efficiency of mobility services and detecting disruptions to traffic or infrastructure. However, the increased connectivity/monitoring introduces privacy and safety risks. Experimental research on those topics would add great value to the current transportation literature.

Bibliography

Ahuja, R.K., Magnanti, T.L., Orlin, J.B., 1993. Network Flows: Theory, Algorithms, and Applications. Prentice Hall.

Bahat, O., Bekhor, S., 2016. Incorporating ridesharing in the static traffic assignment model. Networks and Spatial Economics 16, 1125–1149.

Bai, J., So, K., Tang, C., Chen, X., Wang, H., 2019. Coordinating supply and demand on an on-demand service platform with impatient customers. Manufacturing & Service Operations Management 21, 556–570.

Ban, X., Dessouky, M., Pang, J., Fan, R., 2019. A general equilibrium model for transportation systems with e-hailing services and flow congestion. Transportation Research Part B: Methodological 129, 273–304.

Banerjee, S., Freund, D., Lykouris, T., 2017. Pricing and optimization in shared vehicle systems: An approximation framework. arXiv preprint .

Banerjee, S., Riquelme, C., Johari, R., 2016. Pricing in ride-share platforms: A queueing-theoretic approach. SSRN 2568258 .

Battifarano, M., Qian, Z., 2019. Predicting real-time surge pricing of ride-sourcing companies. Transportation Research Part C: Emerging Technologies 107, 444–462.

Bhat, C.R., 1998. Analysis of travel mode and departure time choice for urban shopping trips. Transportation Research Part B: Methodological 32, 361–371.

Bimpikis, K., Candogan, O., Saban, D., 2019. Spatial pricing in ride-sharing networks. Operations Research 67, 744–769.

Braverman, A., Dai, J.G., Liu, X., Ying, L., 2019. Empty-car routing in ridesharing systems. Operations Research 67, 1437–1452.

Castillo, J., Knoepfle, D., Weyl, G., 2017. Surge pricing solves the wild goose chase, in: Proceedings of the 2017 ACM Conference on Economics and Computation, pp. 241–242.

Chen, H., Zhang, K., Liu, X., Nie, Y.M., 2019. A physical model of street ride-hail. SSRN 3318557 .

Daganzo, C.F., Ouyang, Y., 2019. A general model of demand-responsive transportation services: From taxi to ridesharing to dial-a-ride. Transportation Research Part B: Methodological 126, 213–224.

Di, X., Ban, X.J., 2019. A unified equilibrium framework of new shared mobility systems. Transportation Research Part B: Methodological 129, 50–78.

Di, X., Ma, R., Liu, H., Ban, X.J., 2018. A link-node reformulation of ridesharing user equilibrium with network design. Transportation Research Part B: Methodological 112, 230–255.

Diamond, S., Boyd, S., 2016. CVXPY: A Python-embedded modeling language for convex optimization. The Journal of Machine Learning Research 17, 2909–2913.

Djavadian, S., Chow, J.Y., 2017. An agent-based day-to-day adjustment process for modeling 'Mobility as a Service' with a two-sided flexible transport market. Transportation research part B: methodological 104, 36–57.

Eick, S.G., Massey, W.A., Whitt, W., 1993. The physics of the $M_t/G/\infty$ queue. Operations Research 41, 731–742.

Foley, R.D., 1982. The nonhomogeneous $M/G/\infty$ queue. Opsearch 19, 40–48.

Gössling, S., 2020. Integrating e-scooters in urban transportation: Problems, policies, and the prospect of system change. Transportation Research Part D: Transport and Environment 79.

Hill, R., Griffiths, W., Lim, G., 2018. Principles of Econometrics. John Wiley & Sons.

Lei, C., Jiang, Z., Ouyang, Y., 2019. Path-based dynamic pricing for vehicle allocation in ridesharing systems with fully compliant drivers. Transportation Research Part B: Methodological (forthcoming).

Li, S., Tavafoghi, H., Poolla, K., Varaiya, P., 2019. Regulating TNCs: Should Uber and Lyft set their own rules? Transportation Research Part B: Methodological 129, 193–225.

Lyft, 2019a. Bonuses and Incentives. https://help.lyft.com/hc/en-us/sections/115003494568-Bonuses-and-Incentives.

Lyft, 2019b. New York City Driver Information. https://help.lyft.com/hc/en-us/articles/115012929447-New-York-City-Driver-Information.

Lyft, 2019c. Prime Time for drivers. https://help.lyft.com/hc/en-us/articles/115012926467-Prime-Time-for-drivers.

Ma, H., Fang, F., Parkes, D., 2018. Spatio-temporal pricing for ridesharing platforms. arXiv preprint arXiv:1801.04015 .

Nie, Y.M., 2017. How can the taxi industry survive the tide of ridesourcing? Evidence from Shenzhen, China. Transportation Research Part C: Emerging Technologies 79, 242–256.

Nourinejad, M., Ramezani, M., 2020. Ride-Sourcing modeling and pricing in non-equilibrium two-sided markets. Transportation Research Part B: Methodological 132, 340–357.

NYCTLC, 2019. TLC Trip Record Data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Ozkan, E., Ward, A., 2020. Dynamic matching for real-time ride sharing. Stochastic Systems 10, 29–70.

Perrine, K.A., Levin, M.W., Yahia, C.N., Duell, M., Boyles, S.D., 2019. Implications of traffic signal cybersecurity on potential deliberate traffic disruptions. Transportation Research Part A: Policy and Practice 120, 58–70.

Prékopa, A., 1958. On secondary processes generated by a random point distribution of Poisson type. Annales Univ. Sci. Budapest de Eötvös Nom. Sectio Math 1, 153–170.

Qian, X., Ukkusuri, S.V., 2017. Taxi market equilibrium with third-party hailing service. Transportation Research Part B: Methodological 100, 43–63.

Ramezani, M., Nourinejad, M., 2018. Dynamic modeling and control of taxi services in large-scale urban networks: A macroscopic approach. Transportation Research Part C: Emerging Technologies 94, 203–219.

Rasulkhani, S., Chow, J.Y., 2019. Route-cost-assignment with joint user and operator behavior as a many-to-one stable matching assignment game. Transportation Research Part B: Methodological 124, 60–81.

Rix, K., Demchur, N.J., Zane, D.F., Brown, L.H., 2021. Injury rates per mile of travel for electric scooters versus motor vehicles. The American Journal of Emergency Medicine 40, 166–168.

Rubin, D., 2006. Matched Sampling for Causal Effects. Cambridge University press.

Saleh, W., Farrell, S., 2005. Implications of congestion charging for departure time choice: Work and non-work schedule flexibility. Transportation Research Part A: Policy and Practice 39, 773–791.

Sanders, R.L., Branion-Calles, M., Nelson, T.A., 2020. To scoot or not to scoot: Findings from a recent survey about the benefits and barriers of using E-scooters for riders and non-riders. Transportation Research Part A: Policy and Practice 139, 217–227.

Small, K.A., 1987. A discrete choice model for ordered alternatives. Econometrica: Journal of the Econometric Society , 409–424.

Small, K.A., 1994. Approximate generalized extreme value models of discrete choice. Journal of Econometrics 62, 351–382.

Steed, J.L., Bhat, C.R., 2000. On modeling departure-time choice for home-based social/recreational and shopping trips. Transportation Research Record: Journal of the Transportation Research Board 1706, 152–159.

Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University press.

Wang, H., Yang, H., 2019. Ridesourcing systems: A framework and review. Transportation Research Part B: Methodological 129, 122–155.

Wang, J.P., Ban, X.J., Huang, H.J., 2019. Dynamic ridesharing with variable-ratio charging-compensation scheme for morning commute. Transportation Research Part B: Methodological 122, 390–415.

Wang, X., He, F., Yang, H., Gao, H., 2016. Pricing strategies for a taxi-hailing platform. Transportation Research Part E: Logistics and Transportation Review 93, 212–231.

Wang, X., Yang, H., Zhu, D., 2018. Driver-rider cost-sharing strategies and equilibria in a ridesharing program. Transportation Science 52, 868–881.

Wolsey, L., 1998. Integer Programming. Wiley.

Xu, Z., Yin, Y., Ye, J., 2020. On the supply curve of ride-hailing systems. Transportation Research Part B: Methodological 132, 29–43.

Yahia, C.N., Boyles, S.D., 2021. Peak-load pricing and demand management for ridesourcing platforms, in: 100th Annual Meeting of the Transportation Research Board.

Yahia, C.N., Pandey, V., Boyles, S.D., 2018. Network partitioning algorithms for solving the traffic assignment problem using a decomposition approach. Transportation Research Record 2672, 116–126.

Yahia, C.N., Scott, S.E., Boyles, S.D., Claudel, C.G., 2021a. Unmanned aerial vehicle path planning for traffic estimation and detection of non-recurrent congestion. Transportation Letters .

Yahia, C.N., de Veciana, G., Boyles, S.D., Rahal, J.A., Stecklein, M., 2021b. Book-ahead & supply management for ridesourcing platforms. Transportation Research Part C: Emerging Technologies 130.

Yang, H., Ma, Q., Wang, Z., Cai, Q., Xie, K., Yang, D., 2020. Safety of micro-mobility: analysis of E-Scooter crashes by mining news reports. Accident Analysis & Prevention 143.

Yang, H., Yang, T., 2011. Equilibrium properties of taxi markets with search frictions. Transportation Research Part B: Methodological 45, 696–713.

Zha, L., Yin, Y., Du, Y., 2018a. Surge pricing and labor supply in the ride-sourcing market. Transportation Research Part B: Methodological 117, 708–722.

Zha, L., Yin, Y., Xu, Z., 2018b. Geometric matching and spatial pricing in ride-sourcing markets. Transportation Research Part C: Emerging Technologies 92, 58–75.

Zha, L., Yin, Y., Yang, H., 2016. Economic analysis of ride-sourcing markets. Transportation Research Part C: Emerging Technologies 71, 249–266.

Zhang, K., Chen, H., Yao, S., Xu, L., Ge, J., Liu, X., Nie, M., 2019. An efficiency paradox of uberization. SSRN 3462912 .

Zhang, K., Nie, M., 2019. To pool or not to pool: Equilibrium, pricing and regulation. SSRN 3497808 .

Zhang, R., Pavone, M., 2016. Control of robotic mobility-on-demand systems: A queueing-theoretical perspective. The International Journal of Robotics Research 35, 186–203.

Zou, Z., Younes, H., Erdoğan, S., Wu, J., 2020. Exploratory analysis of real-time e-scooter trip data in Washington, DC. Transportation Research Record: Journal of the Transportation Research Board 2674, 285–299.

Zuniga-Garcia, N., Juri, N.R., Perrine, K.A., Machemehl, R.B., 2021. E-scooters in urban infrastructure: Understanding sidewalk, bike lane, and roadway usage from trajectory data. Case Studies on Transport Policy .

Zuniga-Garcia, N., Tec, M., Scott, J.G., Ruiz-Juri, N., Machemehl, R.B., 2020. Evaluation of ride-sourcing search frictions and driver productivity: A spatial denoising approach. Transportation Research Part C: Emerging Technologies 110, 346–367.