Road Pricing under Heterogeneity*
[preliminary draft; please do not distribute]

Itai Ater† Brad Ross‡ Adi Shany§ Shoshana Vasserman¶

March, 2026

Abstract

Congestion pricing policies are widely debated and critics often argue that toll costs outweigh congestion relief. We show theoretically that welfare gains depend on two forces: driver heterogeneity, which allows tolls to sort high-value drivers onto congested roads, and the extent of revenue recycling. We quantify both forces using data from a large-scale experiment that charged 10,000 Israeli drivers usage-based per-kilometer fees over two years. Reduced-form estimates show that pricing reduced congested driving by roughly 10%, mainly through fewer cross-city trips and shifted departure times, with the strongest responses among drivers with greater schedule flexibility and higher baseline pricing exposure. We then estimate a time-varying traffic equilibrium model with heterogeneous drivers, combining experimentally identified demand with nonparametrically estimated congestion functions. Counterfactual simulations for the Tel Aviv commuting zone show that the experiment's fee schedule would increase welfare by 3% with full revenue recycling but decrease it by 11% without—underscoring that how toll revenues are used is as important as the prices themselves. Optimizing prices increases welfare gains by about 40% under recycling, but optimal prices are zero without it. Cordon pricing generates smaller but more evenly distributed gains.

*First and foremost, we thank Noa Bar-Nir, Fabian Barrett, Hadar Fuks, Shai Gilat, Andrew Meyers, Daniil Mikhailov, Yotam Nir, Doron Zamir, and Michelle Zheng for their exemplary research assistance, as well as the Data, Analytics, and Research Computing team at Stanford GSB for their support. We are also grateful to our colleagues and numerous seminar and conference participants for their helpful comments. This paper subsumes a previous manuscript titled"Can Usage-Based Pricing Reduce Congestion?".

†Tel Aviv University, Coller School of Management and CEPR. Email: ater@post.tau.ac.il.

‡NBER. Email: rossb@nber.org

§Tel Aviv University, Coller School of Management. Email: adishany@tauex.tau.ac.il.

¶Stanford Graduate School of Business and the NBER. Email: svass@stanford.edu.

1 Introduction

Traffic congestion is a notorious source of inefficiency, frustration and environmental damage. On a social level, traffic delays are associated with billions of dollars of waste (Chandler-Wilde, 2023) and substantial greenhouse gas emissions (Barth and Boriboonsomsin, 2008). On an individual level, traffic congestion is associated with serious physical health problems (Currie and Walker, 2011), and time spent in congestion is reported to be among the worst moments of the day (Kahneman et al., 2006). Given these costs, policymakers have sought effective means to reduce congestion, increasingly through policies that charge fees for driving in highly demanded areas. Zone-based entry fees ("Cordon pricing") have been in place in London since 2003, Stockholm since 2006, and Milan since 2008, and were introduced in New York City in 2025.

However, despite the prominence of these examples, congestion pricing policies are often met with heavy resistance. The recent New York City policy came after decades of deliberation and was initially canceled weeks before it was meant to begin in June 2024, only to be abruptly reinstated in November. Although the details underlying the opposition to congestion pricing vary by setting, challenges typically center on the same core argument: a concern that the benefits of diverting some drivers off of congested roads would be outweighed by the monetary costs incurred by drivers who are not diverted and the non-pecuniary costs incurred by drivers who switch to disfavored alternatives.

In this paper, we propose a quantitative framework for assessing this criticism directly. Using a simple illustrative model, we show that the critics have a point: if drivers are homogeneous and congestion fees are not rebated, then any congestion price would be socially wasteful even if congestion levels without pricing are inefficiently high. However, both conditions need to be true for this result to hold. If drivers are heterogeneous—either through preferences, the geography of their trips, or both—then congestion pricing can improve net welfare by incentivizing efficient sorting: diverting drivers with low relative value for using highly demanded roads to benefit drivers with high value. If the distribution of values is sufficiently skewed and congested roads operate at a *steep bottleneck*, then this sorting gain can be sufficiently high so as to outweigh the congestion fees themselves even if the toll revenues are burned upon collection. If, on the other hand, congestion fees are refunded to some degree (either through direct rebates or through valuable public goods) then the speed gains from diverting a fraction of drivers from congested roads can be welfare improving even if all drivers are exactly the same. In reality, both forces likely matter to some degree: drivers are heterogeneous and congestion fees are typically slated for public goods like transit improvements. The question of whether a particular congestion pricing policy is worthwhile is therefore empirical: to what degree can the policy induce efficient sorting, and to what degree do drivers (and other individuals) internalize the benefits of the collected fees?

To demonstrate our framework, we draw on original data from a usage-based congestion pricing experiment that tracked the driving behavior of 10,000 drivers in Israel over the span of two years. Each driver in the experiment installed a monitoring device in their vehicle that could access their GPS coordinates over the course of their trips. During the first six months of participation, participants' driving behavior was monitored, but they did not receive any communication or monetary incentives.

After the initial six-month period, drivers began to incur a time-and-location-specific fee for each kilometer driven out of a virtual budget, the remainder of which was paid out in cash upon drivers' annual vehicle registration renewals.

Our paper is composed of two parts: a reduced form analysis of the experiment and an empirical modeling exercise. In the first part of the paper, we consider the direct effects of congestion pricing on drivers in the experiment: to what extent did treated individuals reduce their driving in congested times and places, and how? To answer these questions, we apply an event study approach in the spirit of Borusyak et al. (2021) to estimate average treatment effects across a number of driving behaviors in the four months after individuals begin to pay for their driving. Our approach relies on two types of variation: variation in when drivers were recruited to the experiment, and variation in drivers' baseline commuting behavior.The dynamic treatment effect estimates are predicated on the staggered recruitment of individuals: an average of about 200 individuals were recruited per week over the course of 50 weeks. As such, on a typical week in which a driver switched from "monitoring" to "treated" status, there was a segment of other drivers who were still in "monitoring" and could serve as a control group to separate individual idiosyncrasies in driving behavior and time-varying aggregate driving conditions from the effects of our congestion pricing treatment. In keeping with Borusyak et al. (2021), our treatment effect estimator models each individual's baseline driving behavior using their "monitoring" period. It then infers the change in behavior due to the onset of pricing by comparing the observed "treated" behavior against a predicted baseline based on the same driver's previous ("monitoring") driving, and the concurrent driving of other drivers still in their "monitoring" phase. To assess the role of heterogeneity, we then apply the Sorted Effects method developed in Chernozhukov et al. (2018) to compare the drivers with the smallest and largest estimated treatment effects across covariates measuring their pre-treatment driving behavior, socioeconomic status, and the accessibility of public transit alternatives to trips that they take by car.

Our dynamic treatment effect analysis suggests that drivers respond immediately, but take several weeks to fully adjust to the congestion pricing scheme. Moreover, after 4 weeks of operating under the pricing treatment, drivers substantially decrease their driving in congested areas and times. Aggregating across time, we find that drivers decrease a variety of measures of congested driving by an average of 10% or more relative to pre-treatment median outcomes, both on the extensive (e.g. number of trips) and the intensive (e.g. departure time) margins. We find substantial heterogeneity across the characteristics of drivers who responded most and least to experimental pricing treatment on both intensive and extensive margins. Using the Sorted Effects method, we find that the most responsive drivers in our sample exhibited more flexibility in their arrival and departure times and took more expensive cross-city and metro-oriented trips during their pre-treatment "monitoring" period. Treatment effects also correlate with the availability of alternatives to driving. Considering an index characterizing the convenience of the best public transit alternative to each vehicle trip taken by car during each driver's "monitoring" period, we find that drivers with the biggest treatment effects on the *extensive* margin (e.g., taking trips altogether) tended to have better access to relevant public transit options. However, public transit access did not correlate strongly with treatment effects on the intensive margin.

In the second part of the paper, we propose a model of supply and demand for traffic throughout the Israeli road network. Under our model, drivers who need to take different trips throughout the road network weigh the convenience, travel time and cost of departing for their trips at different times (or not at all). Depending on which roads their route traverses and what time they depart, each driver intersects with a different number of other drivers taking their own trips. When many drivers load onto the same road segment at the same time, their expected travel time increases according to the "congestion function" of the road segment. In equilibrium, each driver's choice probabilities reflect the expected costs and travel times when all drivers jointly optimize their choices.

A key feature of our model is its focus on heterogeneity in demand for different parts of the road network. Potential drivers are defined at the trip level with an origin, destination, and departure time window and are endowed with mixed-logit preferences over routes, time and money. Routes between different origins and destinations can overlap in many different ways over space and time, and the amount of overlap depends endogenously on equilibrium choices. The congestion effects of overlaps depend on the physical characteristics of the roads (e.g., the width, number of lanes, spacing of exits/entrances etc.) on which they take place. If a pricing policy relieves a bottleneck in one part of the road network, its effects can propagate throughout the rest of the network in non-monotonic ways: some roads might get faster as the trips traversing them reach their destinations sooner, while others might get slower as the formerly bottlenecked drivers make their way through their routes. To capture these complex interactions in a computationally tractable way, we model drivers' positions along their routes using continuous-time Markov chains. This Markov chain representation allows us to efficiently compute the expected number of drivers on each road segment at each point in time over the course of a day given a vector of choice probabilities for each driver. In equilibrium, the expected number of drivers traversing each segment at each time is consistent with choice probabilities that anticipate expected trip durations and prices based on the same traversals.

To take our model to the data, we estimate a mixed logit demand system for drivers in our experiment. Using the subset of trips that drivers take at a regular cadence within a rolling 8-week window, we construct a panel of *trip opportunities* for each driver. For each potential departure time in a trip opportunity, we compute the expected price and duration that the driver would have incurred if they had taken the trip then, given concurrent speeds throughout the road network at that time.[1] We then estimate the preference parameters that best rationalize the shares of departures in each 15 minute interval on each calendar week for each origin-destination pair.

The panel nature of our data has several key features that facilitate identifying driving preferences for a heterogeneous population. First, because we observe individuals making repeated decisions, we are able to attribute substitution patterns to specific combinations of demographics and road-network exposure. Second, because our experiment mechanically shifts prices, each driver in our panel faces exogenous variation in prices that can be used to trace out price sensitivity. Third, to separately identify disutility for travel time from scheduling preferences, we leverage variation in baseline demand for roads over the course of our panel: Israel began enforcing social distancing as early as March 11,

[1] Although our model allows for choices over routes, we find that there is little scope for reducing trip price or duration within our panel. As such, we assume that drivers always use the fastest route based on speed limits for computational efficiency.

2020, but was largely back to normal activity by early 2021. (The Times of Israel, 2021).

Our demand estimates show substantial heterogeneity in value of travel time (VOTT): the median VOTT is about 15% higher than the median hourly wage in Israel, but the 75th percentile is nearly double that. This heterogeneity is correlated with demographics: lower income drivers have a higher disutility from price and a lower sensitivity from travel time. Preferences also vary with the availability of feasible public transportation: drivers with a higher public transit accessibility score are slightly less sensitive to price, and slightly more sensitive to trip duration. In order to accurately capture the interaction between preferences and the road network, we also estimate a supply system for traffic flow. Using high frequency data from sensors along a major highway, we estimate the "fundamental diagram" relationship between traffic speeds and flows on highway segments throughout Israel. We then calibrate population weights to map each trip in our sample to a representative load on the road network using representative origin-destination trip counts from GPS data.

Combining supply and demand, we solve for counterfactual traffic equilibria in which all drivers in Israel's largest commuting zone (the Central District around Tel Aviv) are subject to different congestion pricing schemes. Each simulation equilibrates speeds across 3,700 unique highway segments and 32 fifteen-minute intervals during evening commuting hours (2PM-10PM). In each case, we evaluate the equilibrium distribution of driving behavior, road speeds, and welfare. We begin with a baseline in which traffic is driven by our supply and demand parameters but no congestion prices apply. We then consider an alternative in which *all* drivers are subject to the price schedule set in the experiment. We find that the experiment's price schedule would reduce average driving probabilities by nearly 10%—very close to the extensive margin treatment effect in our experiment and the drop in traffic reported in the immediate aftermath of congestion pricing in London and New York City. However, the gains in speed would be not be enough to offset the toll burden without any revenue recycling.

We then turn to optimal pricing. Using the pricing zones established in the experiment, we optimize over price levels with respect to a utilitarian welfare objective with full rebating. We find that the optimal usage-based prices lower the price for peak-hour driving inside major cities by 21% but raise the price for kilometers driven in off-peak hours and along the metropolitan fringe. Considering a cordon-based alternative, we find that the optimal prices predominately charge trips transitioning from the metropolitan fringe with fees over twice as large during peak hours as in off-peak. The optimal price schedules highlight the distributional effects of targeting on a heterogeneous road network. The optimal usage based prices increase rebate-inclusive welfare substantially for drivers with the highest and lowest price sensitivity but decrease welfare for drivers in the middle. The optimal cordon based prices increase welfare modestly for all drivers. Without rebates, however, no price schedule can increase average welfare at all.

In an upcoming version of this paper, we ask how much rebating is necessary to justify congestion pricing in different scenarios. To do this, we compute a threshold for the share of toll revenues that need to be rebated in order for optimal pricing (usage-based or cordon-based) to induce a net average welfare improvement. We then characterize how equilibrium outcomes under the optimal price schedule change with different levels of rebating. Finally, we consider how the degree of heterogeneity in preferences and network structure affects these results.

1.1 Existing Literature

Pigou (1912) and Knight (1924) first and subsequent theoretical contributions by Vickrey (1969) and Arnott et al. (1993) show that congestion pricing policies can be efficiency-enhancing solutions to reduce traffic congestion. However, empirical research that examines to what extent drivers actually respond to monetary incentives to avoid congested roads, and how overall congestion and welfare may be impacted is limited. Due to data limitations, existing studies often rely on measuring aggregate effects of congestion charges on overall traffic using time variation for identification. These studies often struggle to identify the causal impact of congestion charges since factors other than the implementation of new charges typically also change simultaneously.

In a few cases, studies have used field experiments to examine the impact of congestion pricing. These experiments allow researchers to examine how drivers respond to congestion pricing, and which trips are more sensitive at the individual-driver level. The evidence from these experiments has thus far offered limited support for the effectiveness of congestion pricing. Martin and Thornton (2017) analyze the effects of a congestion pricing experiment in Melbourne, Australia. The experiment collected time and location data from 1,400 vehicles over nine months, during which participants were exposed to one of several congestion pricing treatments and a system of credit accounts similar to ours. A main finding by Martin and Thornton (2017) is that higher prices do not affect work commutes, the use of highly congested roads, or time spent driving at low speeds. They find that drivers adjust to higher prices primarily by taking fewer trips to malls and shops. Kreindler (2024) examines the effect of a field experiment in Bangalore, India. Using detailed travel data on 497 commuters and a usage-based pricing rule, Kreindler (2024) finds that commuters do not change the number of trips that they take, and make only minor changes to their routes otherwise. Extrapolating these results with an equilibrium model, Kreindler (2024) concludes that congestion pricing would not substantially reduce congestion in a city like Bangalore.

Other notable recent contributions include Barwick, Li, Waxman, Wu, and Xia (2024) who use data from household travel diaries and housing transaction data from Beijing. They estimate an equilibrium sorting model of housing location and commuting mode choice with endogenous traffic congestion to identify residents' home and work locations and evaluate the impact of alternative urban transportation policies. Durrmeyer and Martinez (2022) use decennial survey data on commuting choices from Paris to estimate a model that considers individual transportation decisions, equilibrium road traffic levels, and speeds inside a city under different driving restrictions. Almagro, Barbieri, Castillo, Hickok, and Salz (2024) use a combination of survey data, GPS pings, and taxi and public transit records from Chicago to model the welfare frontier of single-price per-mile pricing and investments in public transit frequency. Cook and Li (2025) use traffic data and toll transactions from priced "Lexus" lanes in Seattle to estimate a model of time-varying traffic equilibria and study the welfare trade-off between preserving a lane with guaranteed faster travel and restricting the lanes available for free. Hahn et al. (2023) analyze a field experiment giving subsidies for off-peak public transit use in California and find evidence both that a policy of this type may increase overall welfare, and that a partial equilibrium assessment of welfare may overpredict the potential for welfare gains.

2 Motivating Example

In this section, we present a minimal illustrative example to demonstrate the roles of driver heterogeneity and the marginal value of public funds in determining the optimal level of congestion pricing.

Setup A unit mass of potential drivers each consider whether or not to take a trip along a common road. If they take the trip, each driver i anticipates a utility of:

$$u_i = \underbrace{V_i - \gamma_i D(q) - \alpha_i p}_{\delta_i} + \varepsilon_i,$$

where $D(q)$ is the trip duration when a q share of drivers use the road, p is a toll price for using the road, γ_i is the marginal disutility of trip duration, α_i is the marginal disutility of money, and ε_i is a stochastic idiosyncratic taste shock. Each driver compares their value from driving against an outside option, whose utility we normalize to 0 plus another taste shock ε_{i0}. The probability that driver i takes the trip q_i can thus be written as the probability that the deterministic value of driving δ_i is greater than the difference in stochastic taste shocks $\Delta\varepsilon_i \equiv \varepsilon_{i0} - \varepsilon_i$:

$$q_i \equiv \Pr(\delta_i + \varepsilon_i \geq \varepsilon_{i0}) = \Pr(\Delta\varepsilon_i \leq \delta_i) = F(\delta_i),$$

where F is the CDF of $\Delta\varepsilon_i$. When the taste shocks are distributed i.i.d. according to the Type 1 Extreme Value distribution, $F(\delta_i)$ is given by the standard logit formula: $F(\delta_i) = \exp(\delta_i)/(1+\exp(\delta_i))$. In equilibrium, the total traffic on the road is given by the total expected road use across all drivers:

$$q^* = \sum_i F(V_i - \gamma_i D(q^*) - \alpha_i p).$$

Welfare Objective A utilitarian planner can set the toll p in order to maximize the aggregate money-metric welfare of all potential drivers in equilibrium:

$$W_\Lambda(p) = \sum_i \frac{1}{\alpha_i}\mathcal{W}_i(\delta_i) + \Lambda\, p\, q = \sum_i \frac{1}{\alpha_i}\Big[\underbrace{q_i\,\delta_i}_{\substack{\text{expected}\\ \text{det. payoff}}} + \underbrace{\mathcal{K}(q_i)}_{\text{option value}}\Big] + \underbrace{\Lambda\, p\, q}_{\text{rebated revenue}},$$

where the expected utility from driving for each driver i is the product of the probability that they drive and the expected value of their utility from driving given aggregate traffic level q:

$$\mathcal{W}_i(\delta_i) = \mathbb{E}[(\delta_i - \Delta\varepsilon_i)\cdot \mathbf{1}\{\Delta\varepsilon_i < \delta_i\}] = \delta_i \underbrace{\Pr(\Delta\varepsilon_i < \delta_i)}_{=\,q_i} - \underbrace{\mathbb{E}[\Delta\varepsilon_i \cdot \mathbf{1}\{\Delta\varepsilon_i < \delta_i\}]}_{-\mathcal{K}_i(q)}.$$

The welfare function $W_\Lambda(p)$ sums the expected utility of all drivers in units of money (e.g., weighted by their marginal utility of money $\frac{1}{\alpha_i}$) and adds back the expected revenue from the toll, weighted by the revenue recycling parameter Λ. When $\Lambda = 1$, the toll is fully rebated and each potential driver receives an equal share of the toll revenue regardless of whether or not they drive. When $\Lambda = 0$, the

toll revenue is kept entirely by the planner and each driver bears the full cost of the toll with no rebate.

Efficiency Gap with Homogeneous Types When all drivers have the same preferences (up to their idiosyncratic taste shocks) and tolls are fully rebated ($\Lambda = 1$), the welfare function simplifies to:

$$W_{\Lambda=1}(p) = \frac{1}{\alpha}\big[\, q(p)\,(V - \gamma D(q(p))) + \mathcal{K}(q(p))\,\big].$$

Under this setting, we can recover the classical result that a toll can improve welfare by preventing inefficiently high traffic levels (Pigou, 1920). To see this, note that in equilibrium, the marginal driver must be indifferent between driving and not driving. Because drivers are differentiated only by their taste shocks, this implies that when q is in equilibrium, the deterministic value of driving must be equal to the q^{th} quantile of relative outside option shocks:

$$V - \gamma D(q^*(p)) - \alpha p = F^{-1}(q^*(p)). \tag{1}$$

Comparing this to the first order condition of the welfare function below, it is clear that the equilibrium traffic level without a toll $q^*(0)$ is above the welfare maximizing level q^{opt} whenever $D'(q^{\text{opt}}) > 0$:

$$V - \gamma\big[\, D(q^{\text{opt}}) + q^{\text{opt}} \cdot D'(q^{\text{opt}})\,\big] = F^{-1}(q^{\text{opt}}).$$

However, when $p = (\gamma/\alpha) \cdot q^{\text{opt}} \cdot D'(q^{\text{opt}})$, the two equations are identical. That is, when the toll is equal to the marginal cost of congestion, the optimal level of traffic q^{opt} is attained in equilibrium.

The Role of Rebates The familiar result above relies crucially on the assumption that the toll revenue is fully rebated to drivers. At the opposite extreme where the toll revenue is completely discarded ($\Lambda = 0$), there is *no* positive toll that can improve welfare for a homogeneous population. To see this, suppose that a toll price $p > 0$ could improve welfare. In order for p to be welfare improving relative to not having a toll, it must be that $q^*(p)$ is lower than $q^*(0)$—otherwise, congestion is the same or worse and there is nothing to offset the toll cost. However, in order for $q^*(p)$ to construe an equilibrium, the marginal driver must be indifferent between driving and not driving: $V - \gamma D(q^*(p)) - \alpha p = F^{-1}(q^*(p))$. Since $q^*(p) < q^*(0)$ and $F^{-1}(\cdot)$ is strictly increasing, it must be that $F^{-1}(q^*(p)) < F^{-1}(q^*(0))$ as well. But welfare is strictly increasing in the value of driving and so it decreases when $F^{-1}(q^*(p))$ does—a contradiction.[2]

Note that this argument does not depend on the convexity of the congestion function $D(\cdot)$ or the distribution of the taste shocks $F(\cdot)$. For any toll $p > 0$, we can subtract the equilibrium indifference

[2] We can also show this analytically: equilibrium requires $\delta = V - \gamma D(F(\delta)) - \alpha p$. Implicit differentiation gives

$$\frac{\mathrm{d}\delta}{\mathrm{d}p} = -\frac{\alpha}{1 + \gamma\, D'(q)\, F'(\delta)} < 0,$$

for any increasing trip duration function $D(\cdot)$ since $\alpha > 0$ and $F'(\cdot) > 0$ is a PDF. Thus $\delta(p)$ is strictly decreasing in p. Because $W_{\Lambda=0}(p) = \frac{1}{\alpha}\mathcal{W}(\delta(p))$ and $\mathcal{W}$ is strictly increasing in δ, it follows that $W_{\Lambda=0}(p)$ is strictly decreasing in p.

conditions in Eq. (1) to parametrize the congestion relief that can be sustained in equilibrium as:

$$\underbrace{\gamma\big[D(q^*(0)) - D(q^*(p))\big]}_{\text{congestion relief}} \;=\; \alpha p - \underbrace{\big[F^{-1}(q^*(0)) - F^{-1}(q^*(p))\big]}_{\text{selectivity tightening} > 0}.$$

Since $F^{-1}(q^*(p)) < F^{-1}(q^*(0))$, the selectivity-tightening term is strictly positive, so congestion relief is strictly less than the toll: $\gamma[D(q^*(0)) - D(q^*(p))] < \alpha\, p$. Every inframarginal driver who remains on the road pays $\alpha\, p$ more in tolls but saves only $\gamma[D(q^*(0)) - D(q^*(p))]$ in travel time, suffering a net loss equal to the selectivity-tightening residual. Meanwhile, marginal drivers who exit were indifferent and lose nothing. So all drivers are weakly worse off, and those who stay are strictly worse off.

The Role of Heterogeneity When drivers have heterogeneous preferences, a toll can improve welfare even when $\Lambda = 0$. To see this, suppose that drivers share a common time sensitivity γ but differ in their price sensitivity α_i. When $p = 0$, the α_i heterogeneity is irrelevant: every type faces the same deterministic payoff $V - \gamma D(q)$ and enters at the same rate $q^*(0) = F(V - \gamma D(q^*(0)))$. But once a toll is imposed, types separate. A driver with high α_i (high price sensitivity, low willingness to pay) reduces their driving more than a driver with low α_i. The toll thus reallocates trips away from those who value them least in money terms and toward those who value them most.

To formalize this, decompose welfare at any toll p into three components:

$$W_{\Lambda=0}(p) = \underbrace{G(q)}_{\substack{\text{uniform-entry}\\ \text{benchmark}}} + \underbrace{S(\mathbf{q})}_{\substack{\text{sorting}\\ \text{gain}}} - \underbrace{p\,q}_{\substack{\text{toll}\\ \text{burden}}},$$

where $G(q) \equiv \bar{\eta}\big[q(V - \gamma D(q)) + \mathcal{K}(q)\big]$ is the welfare that would obtain if all types entered at the common rate q (with $\bar{\eta} = \sum_i \omega_i/\alpha_i$ the average marginal value of money), and $S(\mathbf{q})$ captures the extra surplus from heterogeneous entry rates. The sorting term takes the form

$$S(\mathbf{q}) = (V - \gamma D(q))\ \mathrm{Cov}_\omega\Big(\tfrac{1}{\alpha_i},\, q_i\Big) + \sum_i \omega_i \tfrac{1}{\alpha_i}\big[\mathcal{K}(q_i) - \mathcal{K}(q)\big].$$

The first component is a weighted covariance between each type's marginal value of money $1/\alpha_i$ and their entry probability q_i. Since low-α_i types have both a high value of money and a high entry rate under a toll, this covariance is positive: trips are disproportionately taken by those with high willingness to pay. The second component reflects the cost of dispersing entry rates away from the mean—by the concavity of $\mathcal{K}$, spreading q_i around q reduces total option value. The sorting gain $S(\mathbf{q})$ is the net of these two forces.

At the baseline $p = 0$, every type enters at the same rate, so $S(\mathbf{q}^0) = 0$ and $W_{\Lambda=0}(0) = G(q^*(0))$. The welfare change from introducing a toll is therefore

$$\Delta W = \underbrace{G(q^*(p)) - G(q^*(0))}_{\text{congestion relief}} + \underbrace{S(\mathbf{q}(p))}_{\text{sorting gain}} - \underbrace{p\,q^*(p)}_{\text{toll burden}}.$$

We showed above that with homogeneous types, congestion relief falls short of the toll—and with no sorting gain to compensate, welfare falls. With heterogeneous types, the sorting gain $S > 0$ provides a new source of welfare improvement. A toll is welfare-improving whenever this sorting gain is large enough to overcome the net of the toll burden and the (negative) congestion-relief term. This is more likely when there is substantial dispersion in α_i, when the net trip value $V - \gamma D(q)$ is high, and when congestion is severe enough that reallocating trips produces meaningful travel-time savings.

Heterogeneity in Network Exposure The sorting mechanism above relies on drivers differing in their price sensitivity α_i. But a similar logic applies even when all drivers share identical preferences (α, γ, V), provided they differ in their *exposure to the road network*. To see this, consider a road with two sequential segments. A mass m_{12} of drivers must traverse both segments to reach their destination, while a mass m_2 of drivers enter only at segment 2. Let p_1 and p_2 denote the toll on each segment, and let traffic on each segment be $q_1 = m_{12}x_{12}$ and $q_2 = m_{12}x_{12} + m_2x_2$, where x_{12} and x_2 are the respective entry probabilities.

The toll p_1 on segment 1 screens some of the m_{12} through-drivers off the road, which reduces congestion on segment 2 for the m_2 drivers who use only that segment—without charging them. This creates a wedge analogous to heterogeneous α_i: even though preferences are identical, the toll's cost falls entirely on through-drivers while its congestion-relief benefits are shared with segment-2-only drivers who bear none of the toll burden. A welfare improvement requires the congestion function $D_2(\cdot)$ on segment 2 to be sufficiently convex and the value of time γ/α to be sufficiently high, so that the travel-time savings enjoyed by the m_2 segment-2 entrants exceed the toll burden borne by the m_{12} segment-1 entrants. More generally, the potential for welfare-improving tolls without rebates grows as the network creates more degrees of heterogeneity in drivers' exposure to tolled and congested segments.

Connection to our exercise In the remainder of our paper, we quantify the scope for welfare-improving congestion prices along the Israeli road network. We show that drivers exhibit substantial heterogeneity in price sensitivity and time preferences. We then expand our motivating example into a full model of network equilibrium with heterogeneous drivers and use it to assess the level of rebating (or equivalently, the efficiency of government spending) necessary to make congestion tolls welfare-improving.

3 Data

3.1 The Experiment

The primary source of data for this paper is a large-scale field experiment structured toward assessing the impact of "usage-based" congestion pricing, in which drivers pay a per-kilometer fee (similar to a "Vehicle Miles Traveled" (VMT) fee) that varies by time and location. The experiment ran between November 2019 and November 2021 under the administration of Ayalon Highways Co. and

was funded by the Israeli Ministries of Finance and Transportation.[3] Since late 2019, we have served as an independent research team collaborating with the Israeli government regarding the evaluation of the experiment. We joined the experiment after its design was determined.

Participating drivers were recruited in a staggered fashion across the months between January 2020 and October 2020. Approximately 10,000 drivers were recruited in total from across Israel to participate in the experiment, with a mean (median) of 648 (555) drivers recruited each month. Recruitment was done based on pre-defined geographic quotas, designed to capture the relative contribution to congestion from the driving behavior of each area. The majority of drivers reside in the broader Tel Aviv commuting zone, which is home to nearly half of the Israeli population.

To operate the 'driver side' of the experiment, two companies were contracted: (1) Pango, Israel's leading parking payment company, and (2) Pointer, an Auto-Tech company that specializes in the protection and location of stolen vehicles. These companies recruited drivers directly through online marketing and telemarketing campaigns, as well as by approaching existing customers and presenting the experiment to them. Upon recruitment, drivers underwent a brief telephone interview, during which participants were asked to provide administrative information (like their bank account, car license, and car ownership documents) and were given a brief explanation of the experiment's rules. Once participants signed a participation agreement, they had to schedule the installation of a monitoring device, an inconspicuous, difficult-to-remove component that can track when a vehicle's engine is running or idle, as well as its GPS location.[4]

In the first six months after a driver joined the experiment, their driving behavior was monitored and recorded without monetary incentives for any type of driving (we refer to this as the driver's "monitoring" period). After six months in the monitoring period, drivers switched to an "active" status, and began to incur a time- and location-specific fee for each kilometer driven out of a virtual annual budget of 4,500 NIS (approximately $1,250). The virtual budget translated to real money: at each annual registration renewal, drivers were paid the pro-rated remainder of their budget up to 2,000 NIS by an automatic bank transfer. Subsequently, they continued in the experiment under the same conditions but with a renewed initial budget until the following year.[5]

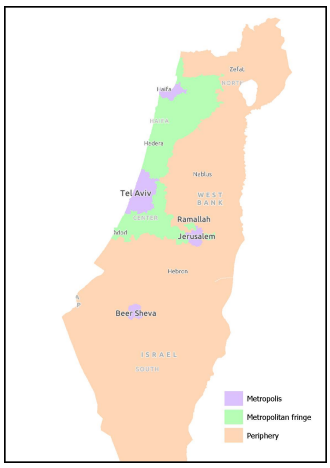
The per-kilometer fee structure was designed to disincentivize driving on the most congested roads. Times were bucketed into "peak" (6:45–9:30am and 3:30–6:30pm), "off-peak" (9:30–3:30am and 6:30–8:00pm) and "leisure" (8:00pm–6:45am), while locations were bucketed into "metro" (the four major Israeli cities, Tel Aviv, Jerusalem, Haifa, and Beer Sheva), "metro fringe" (the commuting zones in which the majority of the Israeli population lives) and "periphery." See Figure 1b for the per-kilometer fees by time of day and geographic zones and Figure 1a for a map of the pricing zones.

Figure 2 summarizes the distribution of drivers' staggered recruitment and activation times. The y axis indexes drivers eventually recruited for our study, and the x axis indexes weeks relative to the first

[3] Ayalon Highways Co. is a metropolitan road company, also responsible for advanced transportation solutions and transportation innovation. It is a government entity similar to a national highway administration.

[4] Most drivers already had a GPS tracking device installed as part of insurance requirements. In such cases, there was no need to install a second one.

[5] The starting annual budget amount was the same in each license renewal period. If the driver's initial active (paid) period was less than 12 months, then their initial budget was reduced proportionally and they started with a full annual budget upon their first license renewal.



(a) Pricing Zones.

	Hours	Metro	Metro Fringe	Periphery
	Weekdays			
Peak	6:45 - 9:30AM 3:30 - 6:30PM	1.5	0.3	0
Off-peak	9:30AM - 3:30PM 6:30 - 8PM	0.1	0	0
Leisure	8PM - 6:45AM	0	0	0
	Weekends			
	All Hours	0	0	0

(b) price in NIS per kilometer traveled

Figure 1: Pricing schedule for the usage-based congestion pricing scheme.

week of recruitment, which was January 5th, 2020. The coordinate corresponding to a given driver and week is colored red if that driver had not yet begun having their driving behavior monitored by that week, yellow if the driver was in their six-month monitoring period during that week, and blue if the driver was in the active, treated state during that week.

For each trip taken by a driver in the experiment, we observe the origin and destination at the level of a Traffic Analysis Zone (TAZ), the distance traveled, the trip duration, and the price that the driver paid (if she was in active status) or would have paid (if she was in monitoring status), based on the per-kilometer fee structure. Similar to US Census blocks, TAZs are geographic polygons that divide the land area of Israel into complete and non-overlapping parcels with relatively homogeneous sizes (in terms of population), but with the primary purpose of analyzing and modeling transit patterns. Due to privacy concerns, we cannot observe the specific location of a driver at any point in time. However, the TAZs can be quite small: metro area TAZs have an average (median) area of 0.7 (0.4) square kilometers (naturally, the periphery area TAZs are bigger), and encode information regarding the types of establishments and land uses (residential, commercial, manufacturing, leisure, etc.) in each TAZ. In addition, while we cannot see the full route that drivers take, we do observe a breakdown of the trip based on the price formula: for each transition between time buckets or geographic zones, we observe the distance, duration, price, and start and end TAZs of that trip segment. Due to data quality issues, we leave incorporating this finer-grained trip information to future work.

In addition to trip characteristics for each driver, we observe each driver's participation status (monitoring or active) on each trip and the date of the status change; the time of payment of the pro-rated remainder budget and the amount paid; whether the dedicated mobile app of the experiment was downloaded after becoming active, as well as how much the driver used the app;[6] and the pollution level of the car and other car characteristics based on its license plate (we could not observe the license plate number itself, for privacy reasons).

[6] A dedicated mobile app was designated the experiment to provide drivers information on the remaining balance of their budget and on estimated time and price for planned trips. drivers were offered to download the app once they switched to the "active" status.

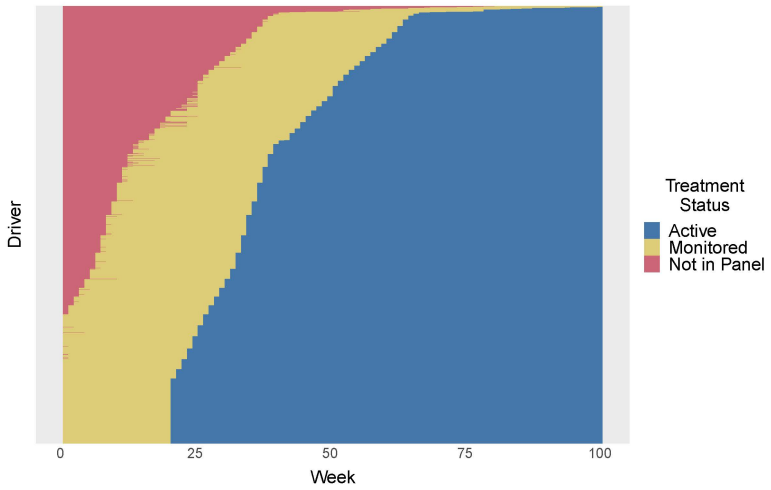


Figure 2: Participant status by week, starting on the week of January 5th, 2020. The y axis indexes drivers eventually recruited for our study, and the x axis indexes weeks relative to the first week of recruitment, which was January 5th, 2020. The coordinate corresponding to a given driver and week is colored red if that driver had not yet begun having their driving behavior monitored by that week, yellow if the driver was in their six month monitoring period during that week, and blue if the driver was in the active, treated state during that week.

3.2 Auxiliary Data

We also supplement our rich data on participants' driving behavior with the following data sources:

Highway sensor data The main highway in the Tel Aviv Metropolitan Area (which is the largest metropolitan area in Israel) is called the Ayalon Highway and is operated and maintained by Netivey Ayalon Co. We received access to traffic volume and traffic speed information collected using Inductive Loop Detectors (ILD) sensors located at different locations along that highway. We use data on average traffic densities and speeds collected at 5-minute intervals over the course of one year to estimate the congestion function mapping traffic densities to traffic flows in Section 8.2.

Open Street Map Since we only observe the start and end TAZ for each trip in our data, we use open-source data on the Israeli road network from Open Street Map to model how different trips would be impacted by changes in road prices and speeds. To do so, we compute the fastest path through the road network between each trip's start and end TAZ, taking road segment traversal speeds to be the speed limits reported in the Open Street Map data. As we find very little variation in the shortest paths computed using time-varying road speeds over the course of our panel, we take the speed-limit based routes as ground truth and consider the impacts of changes in road prices and speeds on driving decisions holding the route fixed.

Public Transit Data We use information on historical public transit schedules and bus stop locations published and made publicly available by the Israeli Ministry of Transportation. We use this information to calculate possible public transit alternatives for the observed trips conducted by the drivers, as described in more detail in Section 4.

Demographics information We use demographic information obtained from the Israeli Central Bureau of Statistics (CBS) for each TAZ that includes information on population density and growth, the population age distribution, the percent of students and employees, and an index of socioeconomic status.[7]

4 Descriptive Statistics

Baseline Driving In our analysis, we include only trips that took place on weekdays (Sunday through Thursday in Israel) and exclude all trips that occur on holidays. Table 1 presents the descriptive statistics of our participants' driving weekly behavior as observed during the "monitoring" (pre-treatment) period. The average (median) driver in our sample drives about 14 (13) times each week and spends 6.2 (5.3) hours behind the wheel each week. The average trip in our sample is 13 Km long. Drivers in the experiment take more trips than the average Israeli driver, who takes 1.9 trips per weekday (about 10 trips per work week). However, their typical trip tends to be shorter in distance and longer in duration, compared to the average Israeli driver who drives 21 Km per trip and spends less than one hour a day on the road.[8] Since the recruitment of the drivers for the experiment was determined based on riders' contribution to congestion, these differences are not surprising.

Table 1 also presents summary statistics broken down by type of trip. We consider three classifications of trips, the sum of which compose all trips: cross-city trips, within-city non-cycle trips and within-city cycle trips. Cycle trips are trips in which the origin and the destination TAZs are the same. Because TAZs in metro areas are quite small, these trips are likely a combination of short-distance trips in less urban places and errand trips more generally. Within-city non-loop trips capture all other driving within the same city. As the table shows, each trip category contributes roughly the same number of trips on average. However, as might be expected, cross-city trips are typically substantially longer, farther and more expensive.

Flexibility covariates To measure the relative flexibility of driving behavior and the relative costs of taking trips at other times, we construct multiple measures using monitoring period data. To characterize drivers' baseline demand for trips around peak hours, we measure the average end time of their last trip in the morning, and the average departure time of their first trip during evening. We focus on the last trip of the morning and first trip of the evening to try to capture the notion of *ideal* timing with regards to work commuting trips (Vickrey, 1969). As Table 2 shows, the average last morning arrival is just before 9:30AM—the endpoint of morning peak hours. The average first evening departure is a bit after 5PM—nearly an hour and a half away from the closest endpoint of evening peak hours. This suggests that if morning and evening trips were equally flexible, the average driver would have a substantially easier time shifting their morning driving to avoid peak hours.

[7] The CBS formulated an index to characterize the local authorities in Israel according to the social level - economic of the population in them. The index is calculated based on several demographic and economic variables, including dependency ratio, average years of schooling in the adult population, percentage of academic degree holders, employment and income levels, etc. Lower values correspond to lower socio-economic status.

[8] Statistics on the general population is based on the most recent available cellular survey data.

Variable	Mean	SD	Quantiles				
			10	25	50	75	90
Panel A: All Trips							
Weekly Trips	14.23	9.45	2.00	7.00	13.00	20.00	27.00
Weekly Km Driven	188.42	190.34	8.00	52.50	137.90	262.90	427.30
Weekly Trips Duration (Hours)	6.23	5.38	0.58	2.47	5.28	8.64	12.49
Weekly Price	46.33	52.93	0.00	4.37	27.68	70.16	123.36
Price / Trip	3.24	3.71	0.00	0.41	2.08	4.74	8.13
Panel B: Cross-City Trips							
Weekly Trips	4.84	5.00	0.00	0.00	4.00	8.00	12.00
Weekly Km Driven	112.06	146.60	0.00	0.00	61.30	166.70	296.81
Weekly Trips Duration (Hours)	3.04	3.64	0.00	0.00	1.92	4.74	7.72
Weekly Price	30.68	45.77	0.00	0.00	8.82	44.56	95.76
Price / Trip	4.24	5.46	0.00	0.00	1.87	6.99	12.44
Panel C: Within-City (Non-Cycle) Trips							
Weekly Trips	4.36	5.46	0.00	0.00	2.00	6.00	12.00
Weekly Km Driven	19.66	34.44	0.00	0.00	7.50	25.00	54.40
Weekly Trips Duration (Hours)	1.19	1.88	0.00	0.00	0.50	1.61	3.30
Weekly Price	6.39	14.94	0.00	0.00	0.28	5.80	18.97
Price / Trip	1.00	2.03	0.00	0.00	0.08	1.10	3.12
Panel D: Within-City Trips (Cycle) Trips							
Weekly Trips	3.68	4.77	0.00	1.00	2.00	5.00	9.00
Weekly Km Driven	19.11	42.76	0.00	0.00	5.20	20.70	50.65
Weekly Trips Duration (Hours)	1.12	1.96	0.00	0.01	0.48	1.45	2.94
Weekly Price	3.96	10.62	0.00	0.00	0.00	2.56	11.92
Price / Trip	1.03	2.78	0.00	0.00	0.00	0.63	2.95

Table 1: Summary Statistics of Main Outcome Variables, Pre-Treatment

To measure flexibility in these trips, we use the Shannon entropy in the distribution of last morning trip arrival and first evening trip departure times. A lower entropy corresponds to a pattern of trips in which the driver arrives (or departs) at around the same time each day of driving, suggesting that their schedule is more rigid. Conversely, a higher entropy suggests that the driver may have more flexibility in when they can take their most frequent trips. As rows two and four of Table 2 show, first evening departure times in our data exhibit substantially higher entropy than last morning arrivals. This is consistent with the idea that there is less flexibility in when drivers arrive to work than when they can leave. However, entropy varies widely across the population at both times of day, suggesting that there is ample heterogeneity in flexibility across our sample.

Personalized public transit indices To get a better understanding of the availability of travel options available to our subjects besides private vehicles, we construct a personalized public transit index. For each trip taken by our drivers during their monitoring period between an origin and a destination TAZ, we look at whether the trip could be made at that time of day with public transit, using our public transit data described in Section 3.2. If a potential public transit route exists between these two locations, we measure the predicted duration of this route at that time of day and take the ratio of the observed trip duration and the predicted duration of the equivalent public transit trip. We then take the average of these trip level duration ratios to compute the driver-level public transit

			Quantiles				
Covariate	Mean	SD	10	25	50	75	90
Last AM Arrival Time	9.48	1.24	7.73	8.67	9.76	10.46	10.80
Last AM Arrival Entropy	1.19	0.51	0.64	0.87	1.22	1.56	1.79
First PM Departure Time	17.08	0.91	16.11	16.43	16.89	17.55	18.30
First PM Departure Entropy	1.60	0.53	0.87	1.33	1.67	1.98	2.23
Public Transit Access	2.97	8.25	1.22	1.87	2.63	3.58	4.75
Home Block Socioeconomic Rank	0.68	0.90	-0.62	0.15	0.81	1.32	1.73
Home Block Population Density	0.01	0.01	0.00	0.00	0.01	0.01	0.02

Table 2: Summary statistics of Pre-treatment Driver Characteristics

index. If a public transit trip cannot be found for a given origin-destination pair, we set the ratio be equal to zero for that trip.

Higher values of our index reflect the availability of public transit options between a driver's observed origin-destinations with relatively shorter travel times compared to the observed driving duration. The average public transit index across our sample is about 3. If the average driver took only one type of trip, this would correspond to a public transit alternative that is 3 times slower than driving. However, the 25th percentile is 1.87 and the 10th percentile is 1.22, suggesting that a segment of our drivers had competitive transit alternatives to their driving.

Socioeconomic covariates To learn about the demographic characteristics of the drivers in our sample, we match the driver home's TAZ with the demographic information from the CBS described in Section 3.2. To infer each driver's home TAZ, we calculate each driver's most frequent origin TAZ of their first trip of the day and their most frequent destination TAZ of their last trip of the day; whichever occurs more frequently is assumed to be the driver's home TAZ. Although this gives us access to all of the demographic information that the CBS provides at the TAZ level we focus the socioeconomic index and population density. Figure 16c in **??** compares the distribution of socioeconomic status in our sample of drivers to the Israeli population writ large. Our sample, colored in pink, clearly skews to the right of the national distribution in blue, under-representing the bottom quartile and over-representing the top quartile of socioeconomic status. This is, in part, to be expected: our recruitment protocol required drivers to have personal vehicles. However, it suggests that our results cannot speak confidently to the effects on the poorest segment of the population.

5 Average Effects of Usage-Based Pricing

Having detailed our setting and data, we turn to evaluating what average effects our congestion pricing scheme had on the driving behavior of participants in the experiment.

5.1 Identification and Estimation

To estimate the average effects of our congestion pricing scheme on driver behavior over time, we would ideally take advantage of random variation in the treatment statuses across the drivers in our sample

in each week, e.g. random assignment of a group of drivers to a never-treated group or randomizing monitoring period lengths across drivers. However, recall from Section 3.1 that Ayalon Highway Co. designed the study so that drivers entered the study in a staggered fashion over time, and every driver was eventually treated after being monitored for the same number of weeks. As such, the only variation in treatment statuses across drivers in any given week of our study is driven by variation across drivers in recruitment times. Furthermore, the week that each driver entered the study may have depended on a litany of unobserved factors that could be correlated with potential driving behavior, such as the spatial recruitment priorities of Pango and Pointer over time and drivers' idiosyncratic eagernesses to begin participation in the study once recruited.

To circumvent these challenges, we rely on our repeated observations of drivers' pre-treatment behavior across weeks to account for persistent unobserved confounders across drivers and the presence of monitored drivers in our study each week to account for variation in driving behavior due to systematic time-varying shocks like COVID-19 prevalence. In particular, we apply the difference-in-differences framework for treatment effect identification and estimation described in Borusyak et al. (2021) and Callaway and Sant'Anna (2021). For each driver i in our study sampled from some superpopulation of Israeli drivers, we let M_i denote the length of their monitoring period and let C_i denote driver i's congestion pricing scheme activation week. Fixing some measure of weekly driving behavior as an outcome of interest, we let $Y_{it}(c)$ denote driver i's potential outcome in week t were the congestion pricing scheme to become active for i in period c, we let $Y_{it}(0)$ denote driver i's control potential outcome in week t if the congestion pricing scheme were not active for i by week t, and we let $Y_{it} := Y_{it}(C_i \mathbb{1}\{C_i < t\})$ denote driver i's observed outcome in week t. Our estimands of interest can be expressed as functions of *cohort average treatment effects on the treated* (CATTs) τ_{ct}, namely the average difference between treatment and control potential outcomes in week t amongst the *cohort* of drivers who were first treated in period $c \leq t$:

$$\tau_{ct} := \mathbb{E}[Y_{it}(c) - Y_{it}(0) \mid C_i = c].$$

To identify CATTs τ_{ct}, we impose a two-way fixed effects (TWFE) model of control potential outcomes as in Borusyak et al. (2021):

$$\mathbb{E}[Y_{it}(0) \mid \alpha_i] = \alpha_i + \gamma_t,$$

where α_i denote driver-specific fixed effects and γ_t denote week effects. The TWFE model implies the parallel trends assumption, namely that drivers' average outcomes would evolve in parallel across weeks in the absence of treatment. Assuming that the number of drivers treated in each week is non-negligible relative to the large sample size of drivers, our CATTs are identified in a standard cohort-specific difference-in-differences fashion (Callaway and Sant'Anna, 2021; Borusyak et al., 2021).

To translate the difference-in-differences identification intuition into an estimation strategy that will also allow us to study treatment effect heterogeneity in Section 6, we follow the imputation approach of Borusyak et al. (2021). We construct consistent estimates $\hat{\gamma}_t$ of the week effects γ_t and asymptotically unbiased estimates $\hat{\alpha}_i$ of the driver fixed effects α_i by running the following fixed effects

regression on drivers' pre-treatment observed outcomes:

$$(\hat{\alpha}_i, \hat{\gamma}_t) := \underset{\alpha_i, \gamma_t}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{C_i > t, C_i - M_i \geq t\} \left(Y_{it} - [\alpha_i + \gamma_t]\right)^2. \tag{2}$$

Here, N denotes number of drivers in our sample and $T < \infty$ denotes the number of weeks in which we observe drivers' outcomes (Borusyak et al., 2021). We impute asymptotically unbiased potential outcomes $\hat{Y}_{it}(0) := \hat{\alpha}_i + \hat{\gamma}_i$ for each driver i and week $t \geq C_i$. This allows us to generate individual treatment effect estimates $\hat{\tau}_{it} := Y_{it} - \hat{Y}_{it}(0)$. To construct consistent estimates of CATTs τ_{ct}, we average the individual treatment effect estimates across drivers in the same cohort as follows:

$$\hat{\tau}_{ct} := \sum_{i=1}^{N} \mathbb{1}\{C_i = c\} \hat{\tau}_{it} \Big/ \sum_{i=1}^{N} \mathbb{1}\{C_i = c\}.$$

We then aggregate our CATT estimates in different ways to obtain the results in the following subsections. To generate event studies, we aggregate $\hat{\tau}_{ct}$ across cohorts to estimate the average CATT at e periods post-treatment τ_e:[9]

$$\hat{\tau}_e := \sum_{c=1}^{T} \mathbb{P}(C_i = c) \hat{\tau}_{c(c+e)}, \quad e = \underline{e}, \ldots, 0, \ldots \bar{e}.$$

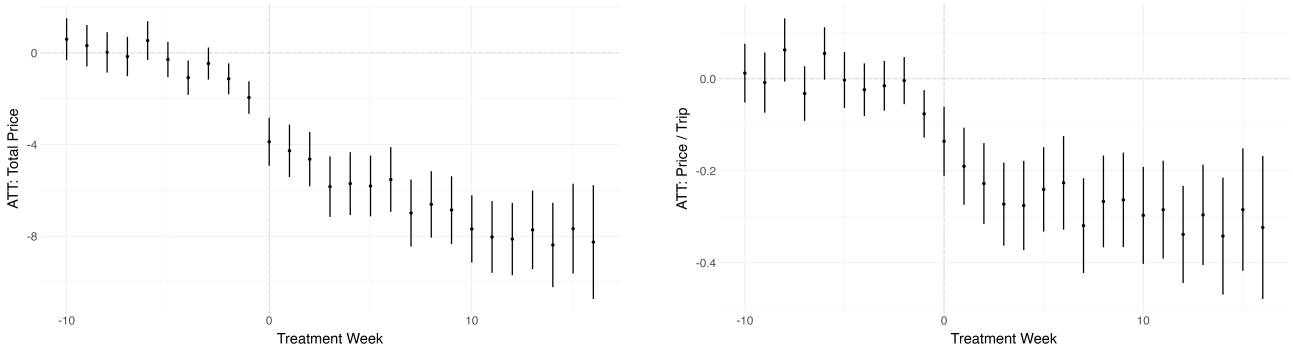
We conduct valid inference on dynamic treatment effect paths by constructing 95% simultaneous confidence bands via a weighted bootstrap procedure akin to the one proposed in Callaway and Sant'Anna (2021) with Dirichlet weights (Shao and Tu, 2012). To obtain average treatment effect estimates across time, we aggregate $\hat{\tau}_e$ over the fourth through 16th weeks relative to treatment:[10] $\hat{\tau} := \frac{1}{12} \sum_{e=4}^{16} \hat{\tau}_e$. We discuss inference for $\hat{\tau}$ in Section 5.3.

5.2 Treatment Effect Dynamics

Figure 3 displays estimates of dynamic average treatment effects on the two most direct payment-related outcomes and associated 95% simultaneous confidence bands over the 10 weeks before and 16 weeks after each driver's treatment week on several outcomes of interest. In Figure 3a, we plot an event study treating the total weekly price deducted from each driver's budget as our outcome of interest. As prices are highest for driving in congested areas and times, a decrease in the total price most directly corresponds to a likely decrease in congestion-inducing behavior. In Figure 3b, we plot a similar event study, using the price per trip as an outcome. The price per trip averages out the total number of trips taken. As such, we interpret effects on it as intensive-margin responses: a

[9] As discussed in Callaway and Sant'Anna (2021), there is a trade-off between estimating effects for larger relative time horizons e and obtaining precise dynamic CATTs. This is because $\tau_{c(c+e)}$ is not estimable for cohorts who are treated for fewer than $c + e$ weeks. In order to avoid contamination due to compositional differences, we estimate τ_e using only the cohorts of drivers for whom we observe outcomes for at least $\bar{e} = 16$ weeks post-treatment.

[10] As we show in Section 5.2, our event study results suggest that participants adjusted their behavioral responses in the first few weeks after activation. In order to focus on the longer-running effects, we exclude the first four weeks in computing aggregate ATTs.

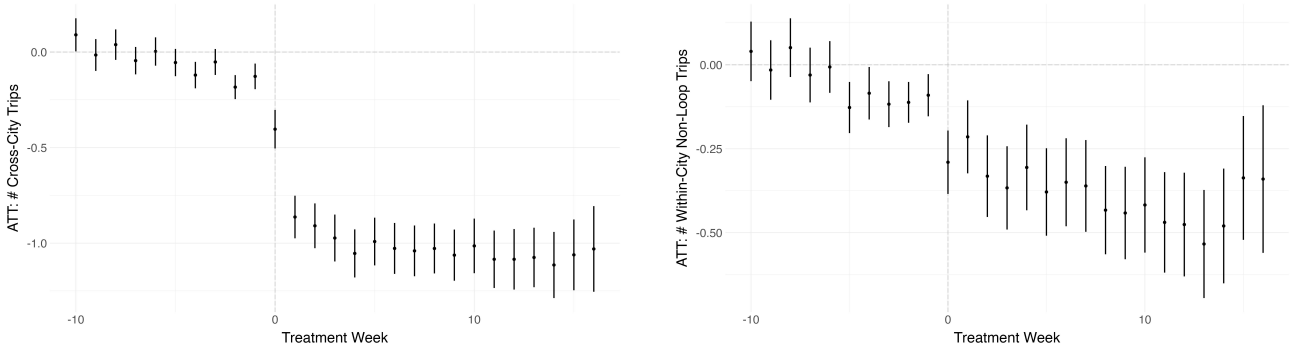


(a) Outcome: total weekly price paid *(b) Outcome: price per-trip*

Figure 3: Event studies of dynamic average treatment effects across weeks relative to treatment week.

lower price-per-trip indicates that treated drivers' trips were less likely to take place in high-priced areas and times on average. The effect on total weekly price seen in Figure 3a is thus driven not only by the numbers of trips that drivers take, but also by the nature of those trips. The magnitudes on both the total weekly price effect Figure 3 are in NIS saved. To put these magnitudes in context, the pre-treatment averages (medians) of these variables are 46 (28) NIS and 3 (2) NIS/trip, respectively.

In addition to the general trend of the dynamic treatment effects, Figure 3 demonstrates two noteworthy features. First, the treatment effects take several weeks to stabilize. This effect lag suggests that drivers may need several weeks to learn how best to adjust their behavior. Second, there appears to be a slight anticipation effect in the 1-2 weeks prior to the start of the treatment period in Figure 3. This is likely due to noise in when drivers were actually alerted that they were going to begin getting treated, as well as to minor inaccuracies in the recording of the exact treatment timing. To assess the robustness of our treatment effect estimates to potential anticipation effects, we find similar results to those plotted in Figure 3 when each driver's effective treatment week is set as two weeks prior to their measured treatment week (Callaway and Sant'Anna, 2021).



(a) Outcome: number of cross-city trips per week *(b) Outcome: number of within-city trips per week*

Figure 4: Event studies of dynamic average treatment effects across weeks relative to treatment week.

To develop a deeper understanding of how drivers adjusted their behavior, in Figure 4 we plot the dynamic average treatment effects on the number of trips taken across two categories: cross-city trips and within-city non-loop (e.g. with different origin and destination points) trips. Figure 4a shows,

the treatment response on cross-city trips was sharp and consistent. By four weeks after treatment, drivers reduced their cross-city driving by about one trip per week on average. Figure 4b also shows a gradual treatment response with an average decrease of 0.25-0.5 trips per week.[11] However, as we show in Appendix A.1, the price-per-trip and total weekly price only exhibit clear effects among cross-city trips. Focusing on cross-city trips, Figure 15 plots the dynamic average treatment effects on departure times for first morning trip and the first evening trip of each day. As Figure 15a shows, drivers not only took fewer cross-city trips after treatment, but they also tended to leave later for the cross-city trips that they took in the evening. In Section 5.3 we show that this difference in departure times also results in less overlap with peak hours. However, this effect seems concentrated in the evening: as Figure 15b demonstrates, there was no clear trend in the treatment effects on the first departure time of cross-city trips taken in the morning.

5.3 Average Treatment Effects Across Outcomes

We now turn to studying the average effects of our congestion pricing treatment on a wider variety of driving outcomes. For each figure in this section, we plot the average treatment effect estimates for a series of outcomes normalized to the units of each outcome's pre-treatment standard deviation. Because the outcomes may be related, we report confidence intervals for each estimate in two ways. In thin lines, we report the Bayesian bootstrap-based confidence interval that controls for the family-wise error rate (FWER) across all of the outcomes in the same plot at 5% (Romano et al., 2010). In thick lines, we report the uncorrected confidence intervals computed with a Bayesian bootstrap that treats each ATT estimate separately. In addition, we color each point estimate and its confidence interval according a hypothesis test against a zero effect. A gray estimate indicates that the stepdown procedure developed in Romano and Wolf (2005) does not reject the null hypothesis of a zero effect size while controlling per-plot FWER at 5%.[12] Correspondingly, a green (*resp.*, red) estimate indicates that the stepdown procedure rejects the null hypothesis of a zero effect size in the direction of an increase (*resp.*, decrease).

Figure 5 plots the average treatment effects on key congestion-related outcomes for cross-city and within-city trips. The ATT on the weekly number of trips in the second row of each panel corresponds to the average of the dynamic CATTs on weeks 4-16 in Figure 4, normalized by the standard deviation of pre-treatment weekly cross-city and within-city trips, respectively. That is, the decrease of approximately 1 cross-city trip per week corresponds to a decrease of approximately 25% of a pre-treatment standard deviation. As Figure 5 shows, the treatment response for cross-city trips is negative in similar relative magnitudes—between 0.1 and 0.2 pre-treatment standard deviations—across other outcomes. This suggests that drivers responded to the pricing treatment not only by reducing their number of cross-city trips, but by specifically reducing their exposure to pricing.

[11] **??** in Appendix A.1 shows that both the number and total price of of within-city loop trips per week stayed about the same on the post-treatment, and even increased a little during the adjustment period.

[12] Note that it is possible for the stepdown procedure developed in Romano and Wolf (2005) to reject a zero effect size in cases where the FWER-corrected confidence interval includes zero. This is because the stepdown procedure is able to be less conservative when rejecting null hypotheses than simultaneous confidence band-based testing procedures while maintaining the same FWER control; to our knowledge, no procedure exists for converting stepdown hypothesis tests into confidence intervals with simultaneous coverage guarantees but shorter widths.

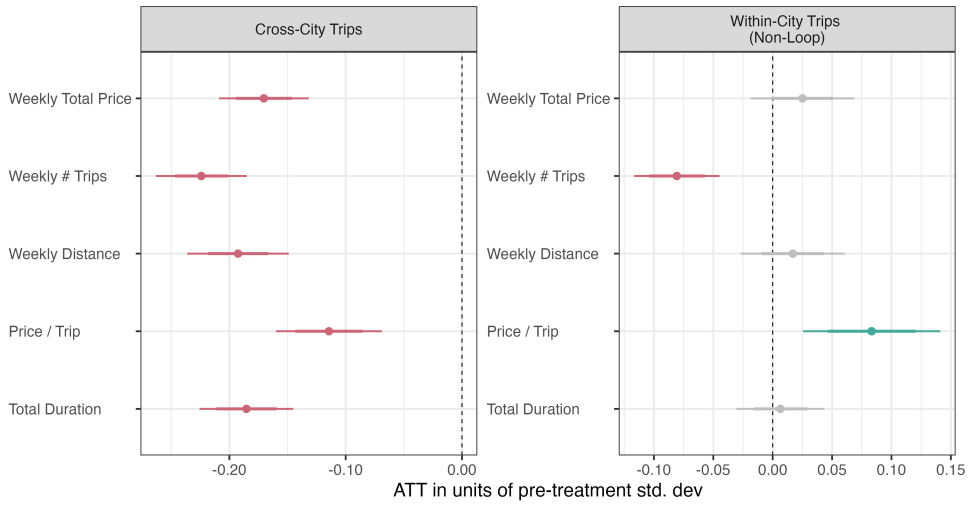


Figure 5: Average Treatment Effects Across Congestion-Related Outcomes

By contrast, there was no substantial impact on pricing-related outcomes for within-city non-loop trips. Total weekly trips decreased by about 7.5%, and the average price per trip increased by about 7.5% but we cannot reject a zero for the impact on total price at 95% confidence. This suggests that driving on some within-city trips may have slightly decreased but the nature of these trips may have increased in exposure to congestion to compensate.

To better understand how drivers adjusted to the pricing treatment, we focus on the two intensive margins of treatment response. Focusing on Cross-City trips, Figure 6 plots the average treatment effects on outcomes related to the time of departure for the first trip of day taken in the morning (left) and the evening (right). As the figure shows, evening trips started about 10% of a pre-treatment standard deviation *later* after the pricing treatment. This corresponds to a decrease of about 10% of a pre-treatment standard deviation in the frequency of trips starting during PM peak hours (and correspondingly an increase in the frequency of PM trips starting during off-peak or leisure hours. The effect on departures for morning cross-city trips was less clear. Although the frequency of trips starting during peak hours seems to decrease somewhat, we cannot reject a zero effect at 95%. Together, these results suggest that drivers responded to the treatment by reducing their exposure to peak hours across the board, but that they may have had more flexibility with respect to their departure times in the evening.

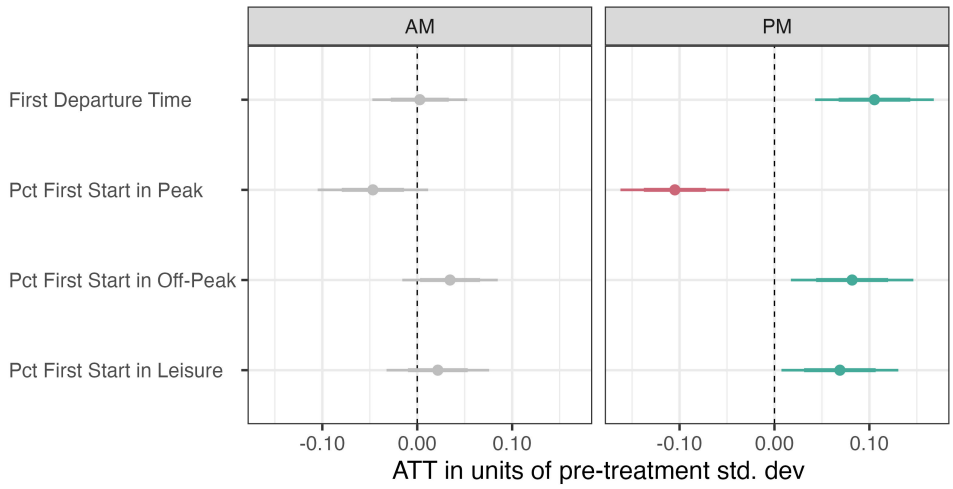


Figure 6: Average Treatment Effects on Departure Times for Cross-City Trips

Figure 7 plots the average treatment effects on the share of each trip that is spent in each of our nine pricing buckets, focusing on cross-city trips. Although none of the treatment effects survives our hypothesis test (with family-wise error correction for all of the other ATTs plotted on the same graph), the point estimates suggest that the share of trip distance covered in metro areas decreased during peak times and increased during leisure times. Similarly, the share of trip distances covered in fringe areas during leisure times shifted up, while the share of trip distances covered in periphery areas during leisure times shifted down. Together, these shifts suggest that the primary way that drivers reduced the price of their trips was by shifting the time in which they drove. This is consistent with the results in Figure 6, which show that drivers did indeed shift their driving time. However, this also suggests that drivers did not respond by changing their routes in a substantive way.

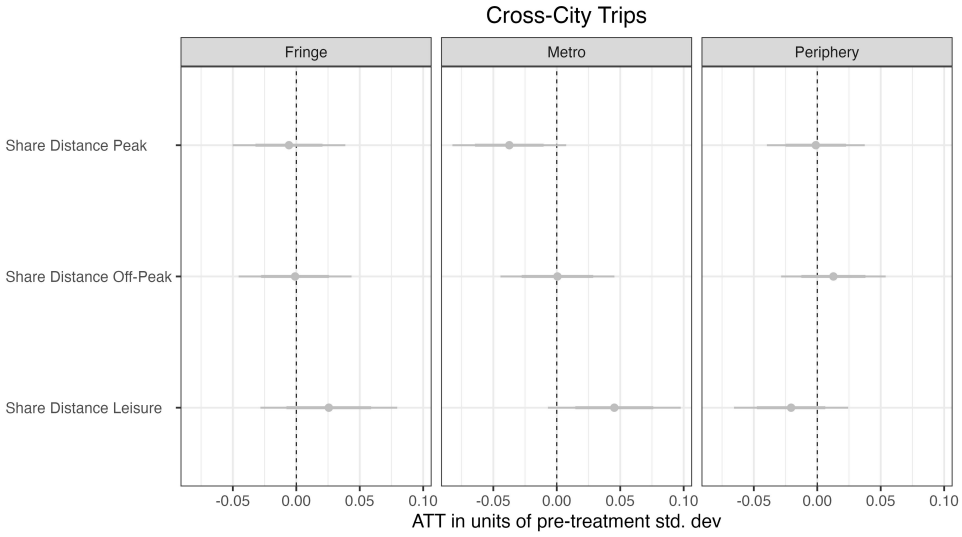


Figure 7: Average Treatment Effects on Trip Shares by Price Bucket for Cross-City Trips

6 Heterogeneity in Responses to Usage-Based Pricing

The control potential outcome model used to identify treatment effects in Section 5 allows for estimating treatment effects at the driver level. Using this variation across drivers, we can test for treatment effect heterogeneity that can be attributed to differences in driver characteristics. To do this, we follow the Sorted Effects method developed in Chernozhukov et al. (2018).

For each outcome of interest, we construct asymptotically unbiased estimates $\hat{\tau}_{it}$ of each driver i's individual treatment effect $\tau_{it} := Y_{it}(C_i) - Y_{it}(0)$ in week $t \geq C_i$ as discussed at the end of Section 5.1. We then consider within-driver averages of $\hat{\tau}_{it}$ across weeks $t = 4, \ldots, 16$, analogous to our average treatment effects in Section 5.3. To test for evidence of statistically significant heterogeneity, we apply the following procedure. First, we project averages of $\hat{\tau}_{it}$ onto a linear combination of rich driver characteristics X_i. This yields predicted driver-specific treatment effect estimates $\hat{\Delta}_i$ and their population analogues Δ_i:[13]

$$\hat{\Delta}_i := X_i \left(\sum_{i=1}^{N} X_i X_i' \right)^{-1} \sum_{i=1}^{N} X_i \cdot \frac{1}{12} \sum_{e=1}^{T} \hat{\tau}_{i(C_i+e)}, \quad \Delta_i := X_i \mathbb{E}[X_i X_i']^{-1} \mathbb{E}\left[X_i \cdot \frac{1}{12} \sum_{e=1}^{T} \tau_{i(C_i+e)} \right].$$

To assess whether the characteristics X_i predict heterogeneous effects across drivers, we test whether the distribution of Δ_i is non-degenerate. We estimate the distribution of Δ_i along a fine grid of quantiles $\{\hat{Q}_{\Delta_i}(u)\}_{u \in [0.05, 0.06, \ldots, 0.95]}$ of the distribution of $\hat{\Delta}_i$ across drivers, and construct 90% uniform confidence bands using a Bayesian bootstrap of the entire estimation procedure from the fixed effects regression (2) through the projection to compute $\hat{\Delta}_i$ (Chernozhukov et al., 2018). We plot the results of this procedure when the outcome of interest is a driver's total weekly price paid in Figure 17 in **??**. Since the uniform confidence band does not contain any horizontal lines,[14] we can reject the null hypothesis that all quantiles of the distribution of Δ_i are the same. As such, we can conclude that there is substantial treatment effect heterogeneity that can be attributed to driver covariates. Analogous results for each of our main outcomes are presented in the online appendix.

6.1 Covariate Differences Between the Least and Most Affected Drivers

We now turn to characterizing the dimensions along which drivers with strong treatment effects differ from those with weak treatment effects. To do so using the Sorted Effects framework, we apply the following procedure. For each outcome of interest, we estimate the vector μ of differences in average characteristics between the 20% of drivers with the weakest (e.g. least negative) covariate-predicted treatment effects and the 20% of drivers with strongest (e.g. most negative) covariate-predicted treatment effects:

$$\mu := \mathbb{E}[X_i \mid \Delta_i \geq Q_{\Delta_i}(0.8)] - \mathbb{E}[X_i \mid \Delta_i \leq Q_{\Delta_i}(0.2)].$$

[13] We let X_i be a fully-interacted quadratic basis for our covariates of interest.

[14] We can see that no horizontal lines pass through the uniform confidence band since the dashed line in Figure 17 corresponding to the upper bound of the confidence interval for smallest quantile $\hat{Q}_{\Delta_i}(0.05)$ of the distribution of Δ_i is below the dashed line corresponding to the lower bound of the confidence interval for the largest quantile $\hat{Q}_{\Delta_i}(0.95)$ of the distribution of Δ_i.

For each figure in this section, we plot our estimates of the covariate differences μ across different outcomes and families of covariate characteristics. In each case, we normalize each covariate difference by the population standard deviation of that covariate to be able to compare differences across characteristics with different units. In addition to each point estimate denoted by a dot, we again use thick lines to denote independent confidence intervals (uncorrected for multiple testing problems) computed via a Bayesian bootstrap of the entire estimation procedure. We use thin lines to denote Bayesian bootstrap-based confidence intervals that control the FWER within each plot at 10% (Romano et al., 2010). A gray color for a point estimate and its corresponding confidence intervals indicates that the stepdown procedure developed in Romano and Wolf (2005) cannot reject the null hypothesis of zero difference while controlling per-plot FWER at 10%. A green (resp., red) color indicates that the drivers with the 20% strongest treatment effects had a significantly higher (resp., lower) average covariate than the drivers with the 20% weakest treatment effects under the step-down procedure.

Building on our discussion in Section 5, Figure 8 plots the sorted effects results for heterogeneity in treatment responses through the weekly number of cross-city trips. The left-most panel considers heterogeneity with respect to the extent to which drivers contribute to congestion in their baseline (pre-treatment) driving. As the left panel shows, the drivers who reduced their cross-city driving the most had 2.5 standard deviations more cross-city trips during the pre-treatment period than the drivers who reduced their cross-city driving the least. They also had more expensive trips—likely because they were longer in distance. All together these comparisons suggest that the drivers who changed their behavior the most on the extensive margin were those who were most exposed to high prices for driving. Because our treatment imposed high prices on driving in congested times and places, these drivers likely also contributed the most to congestion.

In some ways, it is intuitive that drivers who were most exposed to the treatment also responded the most strongly to it. But this is by no means guaranteed ex-ante. Drivers who took the most trips prior to treatment likely also had higher baseline needs for taking these trips. As such, our heterogeneity results suggest that a substantial portion of the drivers with high driving needs also had the flexibility to respond to prices by driving less. The sorted effects results in the middle and right hand panels of Figure 8 provide further backing to this idea. Drivers who reduced their cross-city driving the most had more variability (e.g. higher entropy) in their morning arrival and evening departure times, and had better access to convenient public transit alternatives. On the other hand, there is no discernable difference in socioeconomic status between the most and least responsive drivers.

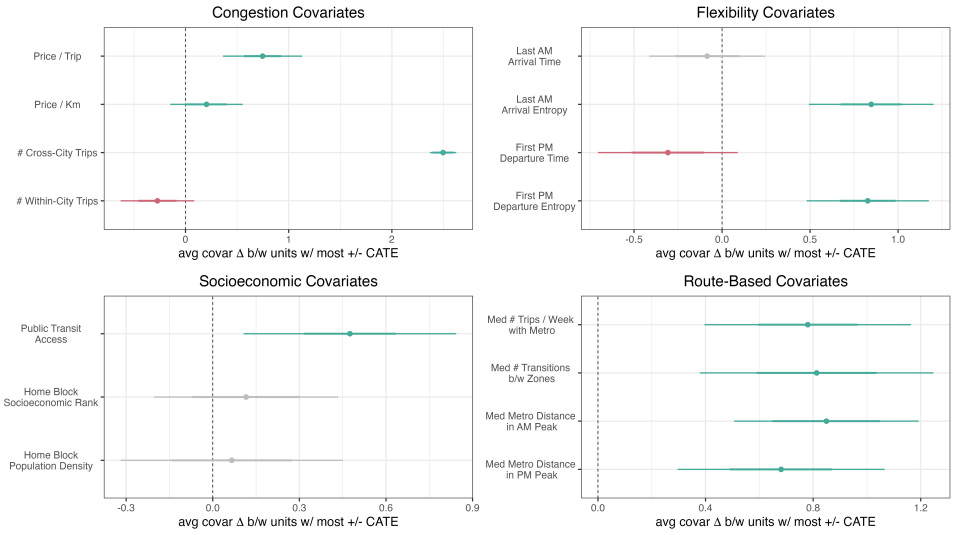


Figure 8: Sorted effects results for treatment effects on the number of cross-city trips taken per week.

Looking to the intensive margin, Figure 9 plots the sorted effects results for treatment responses in the first evening departure time for cross-city trips. As we discussed in Section 5, a stronger treatment response in this context corresponds to a bigger shift to a later first departure time. As with the extensive margin, drivers who responded most strongly with respect to their departure time had higher exposure to the treatment through baseline spending, but they took only 0.5 standard deviations more cross-city trips than the drivers who responded least strongly. Moreover, while measures of flexibility and transit access distinguished drivers with strong extensive margin responses, there was no such distinction for departure time. Rather, the only clear differentiators between drivers with the strongest and weakest effects were their average pre-treatment departure time, the population density of their home TAZ block and the amount of pre-treatment evening driving that took place in metro areas during peak hours. To help interpret these results, it is worth noting that while the average first evening departure time in our sample was about 5:45—a full 45 minutes before the end of peak hours—we are focusing only on the departure time of the *first* trip of each evening. Shifting the first trip to later in the evening implies that all evening driving takes place later as well. As such, our sorted effects results suggest that drivers who changed their evening driving times the most not only had higher exposure to high evening prices, but likely also had the bulk of their expensive driving near the intertemporal price border.

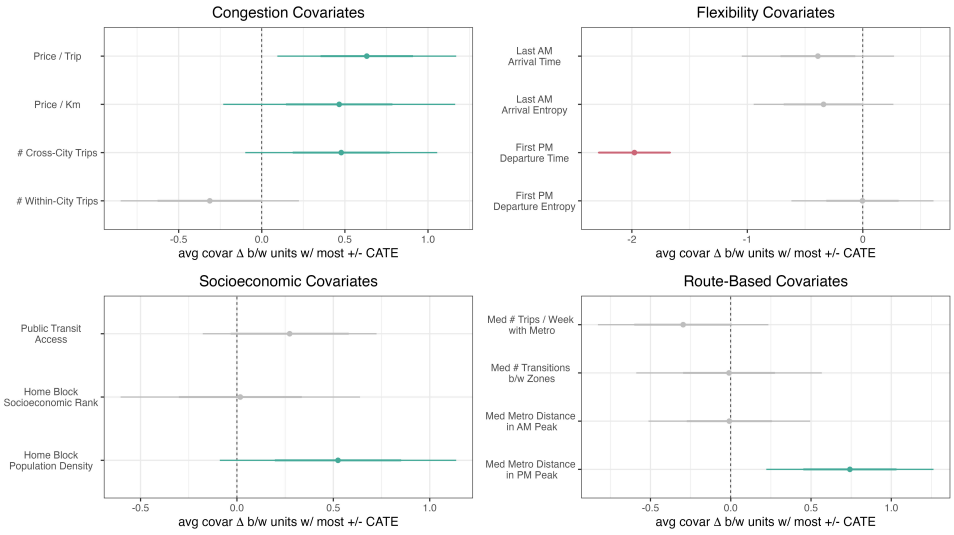


Figure 9: Sorted effects results for treatment effects on the first PM departure time for cross-city trips.

7 A Model of Time-Varying Traffic Equilibrium

We now develop a model of traffic equilibrium along a general road network. Our model builds directly on the simple example presented in Section 2: drivers with heterogeneous values of time consider whether and how to take trips that differ in their origins and destinations given expected costs and travel times. Travel times and trip costs are determined by endogenous traffic speeds throughout the road network, which are themselves determined by drivers' choices of routes and departure times. The resulting congestion dynamics grow in complexity in space and time. When a driver enters the network, the number of other drivers they interact with depends not only on how their different routes intersect, but also on how fast each driver traverses the road segments leading to their meeting points. Our model is designed to capture these rich dynamics on real-world road networks while enabling efficient equilibrium computation and price optimization.

In the model, a mass of Q non-atomic drivers take car trips through a road network comprised of a graph with nodes $n \in \mathcal{N}$ and a set of directed edges $(n, n') \in \mathcal{E} \subseteq \mathcal{N}^2$ corresponding to road segments connecting these nodes. We track drivers' positions in the network in continuous time over a finite interval of length T. We endow each road segment e with a time-varying price function $p_e(\cdot)$, where a driver that finishes traversing segment e at time τ pays price $p_e(\tau)$. This notation nests many canonical road pricing schemes. For example, we can represent usage-based road pricing schemes like the one in the study by setting $p_e(\tau)$ to the per-kilometer price of segment e's pricing zone at time τ times its length, and we can represent cordon pricing schemes like those implemented in London, Stockholm, or New York City by setting $p_e(\tau)$ for segments e that cross a cordon boundary to the price of crossing that boundary at time τ and $p_e(\tau)$ for all other segments to zero.

We let $\mathcal{J}$ denote the set of routes through the network available for drivers to take, where each route $j \in \mathcal{J}$ is characterized by a sequence of road segments $r_j := (e_{j1}, \ldots, e_{jK_j})$ where K_j is the number of road segments in the route, and $e_{jk} \in \mathcal{E}$ denotes which road segment in the road graph

corresponds to the k-th segment in the sequence of road segments comprising route j. We let k_{je} denote the inverse of this indexing scheme, i.e. the index of the road segment in route j that corresponds to road segment $e \in \mathcal{E}$.[15]

Drivers. Driver i chooses whether to make a single trip during the focal time window, and if so, a route J_i to take from their route choice set $\mathcal{J}_i$ and a departure time T_i from their route-specific departure time choice set $\mathcal{T}_{ij} \coloneqq [\underline{T}_{ij}, \overline{T}_{ij}]$.[16] $(J_i, T_i) = (0, \infty)$ denotes a driver choosing their outside option and not taking a trip at all. We assume driver i chooses their route J_i and departure time T_i to maximize their expected utility $U_{ij}(\tau)$ over routes $j \in \mathcal{J}_i$ and departure times $\tau \in \mathcal{T}_{ij}$:

$$U_{ij}(\tau) \coloneqq \underbrace{V_{ij}(\tau) - \gamma_i D_j(\tau) - \alpha_i P_j(\tau)}_{\delta_{ij}(\tau)} + \varepsilon_{ij}(\tau), \quad U_{i0}(\infty) = \varepsilon_{i0}(\infty),$$

where $D_j(\tau)$ and $P_j(\tau)$ are the expected duration and price of a trip on route j departing at time τ to be defined shortly, $V_{ij}(\tau)$ denotes driver i's "scheduling preference" for leaving at time τ when taking route j, and $\varepsilon_{ij\tau}$ are i.i.d. Type-I extreme value shocks over drivers, routes, and times.

Once a driver has departed on route j, they progress along their route according to a continuous-time Markov chain, where upon reaching the k-th segment e_{jk} of route j at time τ, they draw an i.i.d. segment traversal duration D_{ik} from an exponential distribution with rate $\lambda_{e_{jk}:e_{j(k+1)}}$ that we will define shortly and depends on the mass of other drivers also on that road segment at time τ. After D_{ik} time has elapsed, driver i progresses to segment $k+1$ of their route, paying a segment-and-time-specific price $p_e(\tau + D_{ik})$. This route progression process continues until driver i reaches the K_j-th segment of their route and traverses it in time D_{iK_j}. As such, we can define the expected trip duration $D_j(\tau)$ and price $P_j(\tau)$ from departing on route j at time τ as follows:

$$D_j(\tau) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K_j} D_{ik} \,\middle|\, J_i = j,\, T_i = \tau\right], \; P_j(\tau) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K_j} p_{e_{jk}}\left(\tau + \sum_{k'=1}^{k} D_{ik'}\right) \,\middle|\, J_i = j,\, T_i = \tau\right]. \quad (3)$$

While drivers' stochastic traversal times violate the rigid "first-in-first-out" physics emphasized by engineering models of traffic flow,[17] dispensing with this constraint at the driver level allows us to represent spatiotemporally granular traffic dynamics in a computationally tractable manner while remaining faithful to the physics of traffic flow at the population level, as described below.

Segment-level congestion. A canonical model of traffic flow in the transportation engineering literature (see (Boyles et al., 2025) for a textbook treatment) assumes that, at each instant, vehicles' speed $v_e(\rho)$ (measured in distance per unit of time) on a road segment e is a decreasing function

[15] this notation assumes that every route is a simple path in the road graph, i.e. one that traverses each segment in the graph at most once, which is a reasonable assumption.

[16] One can view drivers in this model as single "potential trips" that might be taken, where if a person has the option to take multiple trips during the focal window, each of those trips corresponds to a different driver in the model.

[17] With random traversal times, a driver might traverse a road segment faster than other drivers who entered that same road segment earlier. See Boyles et al. (2025) for a textbook treatment of dynamic traffic assignment models used in civil engineering.

of their density ρ (measured in vehicles per unit of distance) on that road segment, which we take to be uniform along the entire segment. The resulting function relating vehicle density ρ to traffic flow $f_e(\rho) \coloneqq v_e(\rho) \cdot \rho$ through the road segment (measured in vehciles per unit of time) is called the *fundamental diagram* of traffic flow. Typically, the fundamental diagram $f_e(\rho)$ is concave, with flow f_e increasing from zero at zero density up to some critical density $\rho_e^{\text{crit}} \coloneqq \arg\max_\rho f_e(\rho)$, at which point flow decreases until it reaches a jam density ρ_e^{jam}, at which point the rate at which vehicles traverse the segment is zero. We will refer to the fundamental diagram's region of increasing flow as the "free flow" region and the region of decreasing flow as the "congested" region.

Note that the concavity of the fundamental diagram implies that two different vehicle flow rates can be observed for the same vehicle density and speed, which typically leads to multiple equilibria when flow supplied is equated to flow demanded as a function of traversal time (Small and Chu, 2003). To solve this indeterminacy problem, we take a dynamic approach inspired by the canonical Cell Transmission Model (Daganzo, 1994, 1995; Lebacque, 1996), in which we instead assume each road segment is endowed with two different flow functions: a *sending flow* function $f_{Se}(\rho)$, i.e. the rate at which vehicles exit the road segment, as determined by the flow on the "free flow" branch of the fundamental diagram, and a *receiving flow* function $f_{Re}(\rho)$, i.e. the rate at which vehicles can enter the road segment from upstream, as determined by the flow on the "congested" branch of the fundamental diagram:

$$f_{Se}(\rho) = \begin{cases} f_e(\rho) & \rho \leq \rho_e^{\text{crit}} \\ \sup_{\rho'} f_e(\rho') & \rho > \rho_e^{\text{crit}} \end{cases}, \quad f_{Re}(\rho) = \begin{cases} \sup_{\rho'} f_e(\rho') & \rho \leq \rho_e^{\text{crit}} \\ f_e(\rho) & \rho > \rho_e^{\text{crit}} \end{cases}.$$

When density is below ρ_e^{crit}, sending flow is low because fewer vehicles are entering the road segment than the maximum capacity $f_e(\rho_e^{\text{crit}})$ the road segment has to "push" through. When density is high, sending flow should still be high because the road segment can still "push" through as many vehicles as possible up to the maximum capacity $f_e(\rho_e^{\text{crit}})$. On the flip side, when density is low, receiving flow remains high because the road segment has more space to "accept" vehicles than the maximum capacity $f_e(\rho_e^{\text{crit}})$ it has to receive them. When density is high, receiving flow should decrease because vehicles are already packed into the road segment tightly so it's more difficult to accept new vehicles.

Segment-level traversal rates. To explain how this segment-level congestion technology determines drivers' stochastic segment traversal rates, we let $q_{jk}(\tau)$ denote the mass of drivers who chose route j who are currently on the k-th segment of that route at time τ, we collect these time-τ masses into the route-level mass vector $\boldsymbol{q}_j(\tau) \coloneqq \big(q_{j1}(\tau), \ldots, q_{jK_j}(\tau)\big)'$, and we in turn let $\boldsymbol{q}^{\text{route}}(\tau) \coloneqq \{\boldsymbol{q}_j(\tau)\}_{j=1}^J$ denote the collection of all route-level vectors. We then abuse notation slightly and also define the total mass of drivers $q_e(\tau)$ across all routes who are on segment e at time τ as well as the number of drivers $q_{ee'}(\tau)$ who are on segment e at time τ and whose next segments in their routes are e':

$$q_e(\tau) \coloneqq \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e\} q_{jk}(\tau), \quad q_{ee'}(\tau) \coloneqq \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e,\, e_{j(k+1)} = e'\} q_{jk}(\tau). \tag{4}$$

In general, we refer to a pair (e, e') of an upstream segment e and a downstream segment e' as a *movement*, and we let $\boldsymbol{q}(\tau)$ denote the collection of all movement-specific masses $q_{ee'}(\tau)$ at time τ. Since we assume vehicle density is uniform on each segment, we can write the density of vehicles on segment e with length ℓ_e at time τ as $q_e(\tau)/\ell_e$; as such, for notational convenience, we override the definitions of the sending and receiving flow functions to be in terms of vehicle mass instead of vehicle density: $f_{Se}(q) \coloneqq f_{Se}(q/\ell_e)$ and $f_{Re}(q) \coloneqq f_{Re}(q/\ell_e)$

In our multi-route setting, multiple pairs of incoming and outgoing segments could intersect at the same node, meaning not all sending flow from a given incoming segment attempts to enter the same outgoing segment, and the total receiving flow provisioned by an outgoing segment must be divided among the incoming segments. As such, we assume that total sending flow $f_{Se}(q_e(\tau))$ from an upstream segment e is allocated to a downstream segment e' in proportion to the mass of drivers $q_{ee'}(\tau)$ on segment e whose next intended segment is e', meaning we can write the total sending flow of vehicles from segment e to segment e' at time τ is $f_{Se}(q_e(\tau)) \cdot q_{ee'}(\tau)/q_e(\tau)$. To determine how a downstream segment e''s total receiving flow is apportioned to upstream segments' sending flows, we set an explicit priority ranking of the incoming segments $\mathcal{E}_{\to e'}$ to e', where as much of the sending flow from an upstream segment $e \in \mathcal{E}_{\to e'}$ is served by the receiving flow of segment e' until this is none left, at which point any remaining segments in the upstream priority list get zero receiving flow allocated to them. For brevity, we just let $f_{Re'e}(\boldsymbol{q}(\tau))$ denote segment e''s available receiving flow that is allocated to segment e.

Putting these pieces together, for a movement (e, e') and a time τ, we can write the endogenous total rate at which drivers transition from segment e to the next segment e' on their route as the minimum of the proportional sending flow from e toward e' and the priority-allocated receiving flow that e' grants to upstream segment e:

$$\nu_{ee'}(\boldsymbol{q}(\tau)) \coloneqq \min\{f_{Se}(q_e(\tau)) \cdot q_{ee'}(\tau) \,/\, q_e(\tau),\ f_{Re'e}(\boldsymbol{q}(\tau))\}\,,$$

with terminal throughput $\nu_{e:\varnothing}(\boldsymbol{q}(t)) \coloneqq f_{Se}(q_e(t))\, q_{e\varnothing}(t)/q_e(t)$ for drivers completing their trip from the last segment.[18] To translate this overall movement traversal rate into the rate $\lambda_{ee'}$ at which any individual driver moves from segment e to e' introduced above, we divide the total throughput rate $\nu_{ee'}$ by the mass of drivers attempting this transition $q_{ee'}(\tau)$:

$$\lambda_{ee'}(\boldsymbol{q}(\tau)) \coloneqq \begin{cases} \nu_{ee'}(\boldsymbol{q}(\tau)) \,/\, q_{ee'}(\tau), & q_{ee'}(\tau) > 0, \\ \frac{d\nu_{ee'}}{dq}(\boldsymbol{q}(\tau)), & q_{ee'}(\tau) = 0, \end{cases}$$

where when $q_{ee'}(\tau) = 0$, the derivative of $\nu_{ee'}(\boldsymbol{q}(\tau))$ equals the limit of $\nu_{ee'}(\boldsymbol{q}(\tau)) \,/\, q_{ee'}(\tau)$ as $q_{ee'}(\tau)$ goes to zero and equals the free-flow speed of segment e when the sending flow for edge e is binding for the overall throughput $\nu_{ee'}(\boldsymbol{q}(\tau))$.

Route-level mass propagation. Next, we describe how driver mass is propagated through the road network over time. Let $F(\theta)$ denote the joint distribution of drivers' choice sets and preferences

[18] We adopt the convention $0/0 = 0$ for the proportional-allocation terms.

$\theta = (\mathcal{J}, \mathcal{T}, V(\cdot), \alpha, \gamma)$, let $\mathcal{J}(\theta)$ and $[\underline{T}_j(\theta), \overline{T}_j(\theta)]$ denote the route and departure-time-by-route choice sets for a driver of type θ, and let $\delta_j(\tau; \boldsymbol{q}, \theta) \coloneqq V_j(\tau; \theta) - \gamma D_j(\tau; \boldsymbol{q}) - \alpha P_j(\tau; \boldsymbol{q})$ denote the systematic component of utility for a driver of type θ, where we now make the dependence of expected trip duration and price on $\boldsymbol{q}$ explicit. Then we can write the aggregate route-departure-time share density $s_j(\tau; \boldsymbol{q})$ and aggregate trip departure rate $a_j(\tau; \boldsymbol{q})$ of drivers on route j in the following form:

$$s_j(\tau; \boldsymbol{q}) \coloneqq \int_\Theta \frac{\exp(\delta_j(\tau; \boldsymbol{q}, \theta))}{1 + \sum_{j' \in \mathcal{J}(\theta)} \int_{\underline{T}_{j'}(\theta)}^{\overline{T}_{j'}(\theta)} \exp(\delta_{j'}(u; \boldsymbol{q}, \theta)) du} dF(\theta), \quad a_j(\tau; \boldsymbol{q}) \coloneqq Q s_j(\tau; \boldsymbol{q}). \tag{5}$$

Once a driver has departed on route j, the driver-level transition rates $\lambda_{e_{jk}:e_{j(k+1)}}$ define a $(K_j + 1) \times (K_j + 1)$ generator matrix $Q_j(\boldsymbol{q}(\tau))$ for a continuous-time Markov chain representing the driver's position along their route over time, where state $K_j + 1$ is an absorbing state corresponding to trip completion:

$$Q_j(\boldsymbol{q}(\tau))[k+1, k] = \begin{cases} \lambda_{e_{jk}:e_{j(k+1)}}(\boldsymbol{q}(\tau)), & k \in [K_j], \\ \lambda_{e_{jK_j}:\varnothing}(\boldsymbol{q}(\tau)), & k = K_j, \end{cases}$$
$$Q_j(\boldsymbol{q}(\tau))[k, k] = -Q_j(\boldsymbol{q}(\tau))[k+1, k],$$

with all other entries equal to zero. Q_j has a bidiagonal, "birth process" structure reflecting the fact that a driver can only move forward one segment at a time, and the trip completion state $K_j + 1$ has an all-zero column, meaning drivers who complete their trip remain in that state.

The Kolmogorov forward equations then dictate that driver mass propagates along each route j according to the following system of ordinary differential equations (ODEs) across routes:

$$\frac{d}{d\tau} \boldsymbol{q}_j(\tau) = Q_j(\boldsymbol{q}(\tau))\, \boldsymbol{q}_j(\tau) + a_j(\tau; \boldsymbol{q})\, \mathbf{e}_1, \quad \boldsymbol{q}_j(0) = \mathbf{0} \quad j \in \mathcal{J}. \tag{6}$$

The first term propagates existing mass forward along the route at endogenous rates encoded in Q_j: for each segment k, mass flows out at rate $Q_j[k, k]$ and into the next segment via $Q_j[k+1, k]$. The second term injects new mass at rate $a_j(\tau; \boldsymbol{q})$ into the first segment (via the standard basis vector $\mathbf{e}_1$). The initial condition $\boldsymbol{q}_j(0) = \mathbf{0}$ dictates that no drivers are on any route at the start of the horizon.

Expected trip duration and price. Equipped with route-level mass dynamics, we close the model by characterizing the expected duration $D_j(\tau; \boldsymbol{q})$ and price $P_j(\tau; \boldsymbol{q})$ of a trip for a driver who has chosen route j and departed at time τ defined in (3). Since both objects can be written as expectations over the stochastic path generated by the route-j traversal process, they can both be computed as special cases of a backward recursion in process state starting from the trip completion state.

Let $X_{jk}(\tau; \boldsymbol{q})$ denote the expected remaining cumulative quantity (either duration or price) for a driver on route j currently in segment-state k at time τ. Suppose that being in segment k on route j at time τ accrues a running amount $r_{jk}(\tau)$ per unit time, that a jump from segment k to $k+1$ incurs a lump-sum amount $m_{jk}(\tau)$, and that if a driver is still in segment k of route j at the horizon time τ, we assign terminal continuation value $\bar{X}_{jk}$. At the absorbing state there are no further contributions, so $X_{j(K_j+1)}(\tau; \boldsymbol{q}) = 0$.

To derive the ODE governing $X_{jk}(\tau;\boldsymbol{q})$ intuitively, note that over a small time increment $d\tau$, a driver in segment k at time τ either transitions to segment $k{+}1$ (with probability $Q_j(\boldsymbol{q}(\tau))[k{+}1,k]\cdot d\tau$) or remains in segment k (with the complementary probability).[19] In either case, the driver accumulates the running amount $r_{jk}(\tau)\cdot d\tau$ over the interval $[\tau,\tau+d\tau]$. By the law of total expectation:

$$\begin{aligned} X_{jk}(\tau;\boldsymbol{q}) = r_{jk}(\tau)\,d\tau + Q_j(\boldsymbol{q}(\tau))[k{+}1,k]\,d\tau\cdot\big(m_{jk}(\tau)+X_{j(k+1)}(\tau{+}d\tau;\boldsymbol{q})\big) \\ + \big(1-Q_j(\boldsymbol{q}(\tau))[k{+}1,k]\,d\tau\big)\cdot X_{jk}(\tau{+}d\tau;\boldsymbol{q}). \end{aligned}$$

Rearranging and taking $d\tau\to 0$ gives the following generic backward ODE governing $X_{jk}(\tau;\boldsymbol{q})$ for each $k\in[K_j]$:

$$-\frac{d}{d\tau}X_{jk}(\tau;\boldsymbol{q}) = r_{jk}(\tau)+Q_j(\boldsymbol{q}(\tau))[k{+}1,k]\big(m_{jk}(\tau)+X_{j(k+1)}(\tau;\boldsymbol{q})-X_{jk}(\tau;\boldsymbol{q})\big). \tag{7}$$

We solve this ODE backward from the post-horizon boundary condition $X_{jk}(\tau;\boldsymbol{q})=\bar{X}_{jk}$. In words, the expected remaining cumulative quantity from segment k decreases over time at a rate equal to the current running contribution plus the expected change generated by a jump: with rate $Q_j[k+1,k]$, the driver incurs the lump-sum amount m_{jk} and moves to a state whose continuation value differs by $X_{j(k+1)}-X_{jk}$.

Expected remaining trip duration $D_{jk}(\tau;\boldsymbol{q})$ starting from the k-th segment on route j at time τ is a special case of $X_{jk}(\tau;\boldsymbol{q})$ with running amount $r_{jk}(\tau)=1$, no jump amount $m_{jk}(\tau)=0$, and post-horizon value $\bar{X}_{jk}$ equal to the free-flow remaining duration. Similarly, expected remaining price $P_{jk}(\tau;\boldsymbol{q})$ starting from the k-th segment on route j at time τ is the special case with no running amount $r_{jk}(\tau)=0$, jump amount $m_{jk}(\tau)=p_{e_{jk}}(\tau)$, and post-horizon value $\bar{X}_{jk}=0$. Solving these specialized $X_{jk}(\tau;\boldsymbol{q})$ ODEs backward from each time τ yields the objects entering demand: $D_j(\tau;\boldsymbol{q})=D_{j1}(\tau;\boldsymbol{q})$ and $P_j(\tau;\boldsymbol{q})=P_{j1}(\tau;\boldsymbol{q})$.

Equilibrium. Finally, we can define a time-varying traffic equilibrium in our model as a collection of movement-level mass functions $\boldsymbol{q}^*(\cdot)$ such that the solutions to the expected duration and price ODEs specialized from (7) yield route departure rates (5) that in turn imply route-level mass propagation according to (6) that aggregate back to the same movement-level mass functions $\boldsymbol{q}^*(\cdot)$ via (4). We use $G(\boldsymbol{q}(\cdot))=\boldsymbol{q}(\cdot)$ to refer to this fixed point equation that an equilibrium $\boldsymbol{q}^*(\cdot)$ solves.

In words, the fixed point requirement for an equilibrium is at the movement level: movement-level masses determine congestion and therefore transition rates, which in turn determine the backward recursions (expected duration and price), which feed into route and departure time demand, which determine route-level inflows, which pin down route-level mass paths together with the transition rates. Aggregating those route-level paths back to the movement level must reproduce the original movement-level masses.

[19] While it is possible the driver could draw very short traversal durations D_{ik} and therefore traverse multiple segments within any interval $d\tau$, this possibility vanishes as we take the limit as $d\tau$ goes to zero.

Discretization and computation. The model above admits a convenient time discretization that facilitates efficient equilibrium computation in practice; we provide a high-level overview of our approach here and defer a detailed description to Appendix B.

In particular, we discretize the length-T time horizon into discrete periods (τ_t, τ_{t+1}) of length Δ indexed by t and assume that segment traversal rates in each period t remain constant at their values $\lambda_{e:e'}(\bar{\boldsymbol{q}}_t)$ when evaluated at the period-average movement-level masses $\bar{q}_{ee't}$:

$$\bar{q}_{ee't} \coloneqq \Delta^{-1} \int_{\tau_t}^{\tau_{t+1}} q_{ee'}(\tau)\, d\tau.$$

This within-period traversal rate homogeneity allows us to express the solutions of the ODE systems governing mass propagation (6) and expected trip durations and prices (7) as linear equations that propagate start-of-period route-level masses forward in time and start-of-period duration and price continuation values backward in time. $\bar{q}_{ee't}$ can also be expressed conveniently via another linear equation in these start-of-period route-level masses.

Since expected trip durations $D_j(\tau; \bar{\boldsymbol{q}})$ and prices $P_j(\tau; \bar{\boldsymbol{q}})$ are constant with in each interval t, we can write them as $D_{jt}(\bar{\boldsymbol{q}})$ and $P_{jt}(\bar{\boldsymbol{q}})$. If we also assume scheduling preferences $V_j(\tau; \theta)$ for a driver with type θ equal constants $V_{jt}(\theta)$ within each period t and that drivers depart at a uniform rate within each period, we can express the rate $a_{jt}(\bar{q})$ at which drivers depart along route j within each period t in a form analogous to mixed logit demand over routes and departure windows:

$$\begin{aligned}
\delta_{jt}(\bar{\boldsymbol{q}}, \theta) &\coloneqq V_{jt}(\theta) - \gamma D_{jt}(\bar{\boldsymbol{q}}) - \alpha P_{jt}(\bar{\boldsymbol{q}}), \\
a_{jt}(\bar{\boldsymbol{q}}) &\coloneqq \frac{Q}{\Delta} \int_\Theta \frac{\exp(\delta_{jt}(\bar{\boldsymbol{q}}, \theta))}{1 + \sum_{j' \in \mathcal{J}(\theta)} \sum_{t' \in \mathcal{T}_{j'}(\theta)} \exp(\delta_{j't'}(\bar{\boldsymbol{q}}, \theta))}\, dF(\theta).
\end{aligned}$$

We can then define a discretized equilibrium fixed point equation in terms of the period-average movement-level masses $q_{ee't}$ analogous to the continuous-time version G. See Appendix B.1 for details.

While the linear operators defining the linear systems governing discretized equilibrium dynamics do not admit closed forms, we describe how to evaluate them approximately in a computationally efficient manner in Appendix B.2 using a technique called uniformization (Stewart, 1994). We provide a complete description of the accelerated fixed point algorithm we use to solve for discretized equilibria in Appendix B.3. In Appendix B.4, we show how the smoothness of our equilibrium allows us to use first-order methods to optimize different segment pricing schemes to maximize welfare under different levels of revenue rebating.

8 Estimating Demand and Supply for Roads in Israel

In order to take our model to the data, we need to estimate how a representative population of drivers' demand for trips through the road network depends on trip durations and prices, as well as how those trips are propagated through the road network via traffic flow supply. In this section, we outline our approach for each of these components.

8.1 Demand for Trips

In order to identify preferences over driving time, we must first establish when an individual might want to drive in the first place. This raises a fundamental challenge: we only observe trips that are taken by drivers in our sample; not all trips that were considered being taken. To resolve this, we focus on trips that seem to reoccur with some regularity. We define a "habitual trip" as a route between the same origin and destination that the same driver takes within a two hour window at least 3 times in a rolling 8 week period. For each habitual trip, we then consider a series of *trip opportunities* at which the trip might have occurred. For instance, if we observe a driver often driving between a given origin and destination around 6 PM, then we consider the possibility that they would take this trip between 5 PM and 7 PM every weekday until we stop seeing them make that trip at the same cadence.

In order to allow preferences to vary flexibly across travel purposes and times of day, we refer to each consistent trip i as a separate driver for the purposes of our model. On each date $\ell \in \mathcal{L}_i$ in which their trip is defined, driver i considers a set of choices $\{(j,t) : j \in \mathcal{J}_i, t \in \mathcal{T}_{ij}\}$ choices for how to take their trip (or not at all). In keeping with our definition of habitual trips, we allow drivers to choose from 9 departure times in 15-minute intervals within an hour of their median departure time in our data. For simplicity, we assume that the driver always takes the fastest route available based on road-level speed limits regardless of traffic conditions and suppress the route index j in what follows. This simplification is consistent with our reduced form finding that drivers did not appear to change their routes in response to the experiment's pricing scheme.[20] As such, it is unlikely to meaningfully restrict the choice set that each driver faces in our estimation model. However, it may be restrictive for counterfactuals that substantially change the geography of where prices apply.

For each potential departure time t on date ℓ, we impute the expected trip duration $D_{i\ell t}$ and trip price $P_{i\ell t}$ that driver i would incur if they were to drive at that time based on concurrent road speeds along their route on that date/time. To do so, we first infer real-time speeds on every road segment in the Israeli road network at every date/time in our panel using high frequency GPS pings from the Israeli public bus network. We then estimate a trip-level regression model to account for systematic differences between observed trip durations and the durations projected based on concurrent road speeds.[21] The expected duration $D_{i\ell t}$ for each trip opportunity (i,ℓ,t) is given by the regression-adjusted projected travel time for trip i based on the road speeds starting at time t on date ℓ. Because we don't expect drivers to check Waze before deciding when to depart every single day, we assume that drivers make departure time choices based on the quarterly average trip duration at time t, denoted by $\hat{D}_{i\ell t}$. We let $Y_{i\ell t} = 1$ if on date ℓ, driver i chose departure time t and $Y_{i\ell t} = 0$ otherwise.

Our panel of trip opportunities for each driver allows us to estimate mixed logit preferences at a granular level in the vein of a standard demand estimation problem (Berry, 1994; Berry et al., 1995). We define markets $m(i,\ell)$ as tuples of dates $\ell \in \mathcal{L}_i$ and origin-destination pairs. We parameterize the distribution of random duration and price coefficients (γ_i, α_i) as a joint Normal distribution with a mean vector depending linearly on a vector of demographic variables X_i (in particular, drivers'

[20] In addition, we computed the real-time shortest path for a randomly drawn sample of routes, days and times throughout our sample. Across these samples, the average ratio of real-time shortest path travel time to the shortest path based on speed limits was within a few percentage points.

[21] See Appendix B for details.

public transit and home-neighborhood socio-economic status indices described in Section 3) and a general covariance matrix Σ. Given our rich panel of drivers' repeated choices over time, we estimate the parameters of our demand model by maximizing a panel mixed logit maximum likelihood with an added IV-GMM "penalty" akin to the CLEER estimator proposed by Grieco et al. (2025). See Appendix C for details.

We identify the price coefficient distribution using panel variation in whether individual drivers are charged non-zero prices for their trips as part of our experiment. Importantly, market-mean duration could plausibly be positively correlated with drivers' unobserved scheduling preferences, since times when more drivers want to be departing (e.g. rush hour being the earliest drivers can leave work) are also the times when trip durations are the longest due to congestion. To remedy this endogeneity issue, we construct instruments that take advantage of fluctuating COVID-19-induced travel norms during our sample period. In particular, we use public transit trip counts departing from a market origin's city across weeks as a proxy for workers' ability and/or need to be in the office: more people traveling from a driver's origin city via public transit does not directly affect the focal driver's intrinsic preference for driving at that time, but it does proxy for increased traffic on that driver's route on account of more people making trips in person.

Results We report the results of our demand estimation procedure in Table 3. We note that on average, drivers dislike higher prices and travel times (based on their mean taste coefficients), higher income drivers are less price-sensitive and more duration-sensitive, and drivers with more access to public transit are both less price-sensitive and less duration-sensitive. We also find that there is some remaining unobserved heterogeneity in preferences unexplained by demographics. To benchmark our estimates, we find that the median Value of Time γ_i/α_i (measured in NIS per hour) for drivers in our sample based on our estimates is 87 NIS, which is quite close to a reported estimate of the median hourly wage in Israel of around 76.5 NIS (ERI SalaryExpert, 2026) and well within typical relative-to-hourly-wage estimates found in the literature.[22] See Appendix D.3 for the full distribution.

8.2 Supply of Traffic Flow

To estimate the congestion functions f_e that govern how traffic density on road segment e affects the flow at which cars traverse e, we use high-frequency data on contemporaneous measurements of traffic densities and speeds, as described in Section 3. Our data includes several years of contemporaneous road flow and speed observations at 5 minute intervals for all of 2020 and 2021 from several sensors along a segment of the Ayalon, the largest highway in Israel. Given the high frequency of our data, we match sensors to Open Street Map road segments, construct estimates of road density as the ratios of observed flows and speeds, and estimate a fundamental diagram f_e for each bundle of unique highway segment characteristics (e.g. speed limit and number of lanes) nonparametrically using a sieve regression with a Bernstein polynomial basis where we restrict the functional form of our estimate to be concave in keeping with the model in Section 7. We plot an example of one of our estimated congestion

[22] See Section 5.2.2 of Su (2022) for a thorough review of different literatures' estimates of similar quantities.

	Estimate	Std. Error	95% CI
Mean Taste Coefficients			
Price (α_0)	-0.1196^{***}	(0.0026)	$[-0.1246, -0.1146]$
Travel Time (γ_0)	-2.3548^{*}	(1.3621)	$[-5.0245, 0.3148]$
Interactions with Price			
SES × Price	0.0316^{***}	(0.0029)	[0.0259, 0.0373]
Transit Score × Price	0.0121^{***}	(0.0010)	[0.0101, 0.0142]
Interactions with Duration			
SES × Travel Time	-1.4244^{***}	(0.0781)	$[-1.5775, -1.2713]$
Transit Score × Travel Time	-0.5503^{***}	(0.0716)	$[-0.6907, -0.4099]$
Random Coefficient Covariance (Σ)			
σ^2_α	0.0003	(0.0003)	$[-0.0003, 0.0010]$
$\sigma_{\alpha,\gamma}$	0.0248^{***}	(0.0046)	[0.0157, 0.0338]
σ^2_γ	2.2313^{**}	(1.0114)	[0.2490, 4.2137]

$^{***}\,p < 0.01$, $^{**}\,p < 0.05$, $^{*}\,p < 0.10$

Table 3: Demand model estimates

functions in Figure 10. We then assign our estimates of these congestion functions at different sensors to different road segments based on their speed limits and numbers of lanes.

Lastly, we re-weight our panel of drivers' consistent trips to be representative of by ensuring that counts of trips between origins and destinations proportionally match a dataset of origin-destination flows we acquired that was derived from a sample of GPS pings from peoples' cell phones. We then choose the total mass of drivers Q to minimize the error we incur when predicting total Ayalon sensor counts in 6-hour windows using counts of observed consistent trips during those same 6-hour windows.

9 Equilibrium Responses to Road Pricing in Israel

Using our estimates from Section 8, we simulate traffic equilibria under several different pricing schemes. We focus our simulations on trips within the largest commuting zone in Israel, the "Central District" around Tel Aviv. For each pricing scheme, we solve for the movement-level vehicle masses that equilibrate speeds for over 3,700 unique highway road segments across the 32 15-minute intervals between 2 PM and 10 PM, as described in Section 7. We then compute driver-level money-metric equilibrium welfare with ($\Lambda = 1$) and without ($\Lambda = 0$) revenue recycling according to Eq. (8), and compare it to equilibrium welfare without pricing:

$$w_{i\Lambda}\big(\{p_e(\cdot)\}_{e\in\mathcal{E}}\big) \coloneqq \frac{1}{\alpha_i}\log\left[1+\int_{\underline{T}_i}^{\overline{T}_i}\exp\big(\delta_{ij(i)}(\tau;\boldsymbol{q}^*)\big)\,d\tau\right] + \frac{\Lambda}{Q}\int_0^T a_{ij(i)}(\tau;\boldsymbol{q}^*)\,P_{j(i)}(\tau;\boldsymbol{q}^*)\,d\tau, \tag{8}$$

where $j(i)$ denotes driver i's route,[23] $\delta_{ij(i)}(\tau;\boldsymbol{q}^*) \coloneqq V_{ij(i)}(\tau) - \gamma_i D_{j(i)}(\tau;\boldsymbol{q}^*) - \alpha_i P_{j(i)}(\tau;\boldsymbol{q}^*)$, and $\boldsymbol{q}^*(\cdot)$ is an equilibrium, i.e. $G(\boldsymbol{q}^*(\cdot)) = \boldsymbol{q}^*(\cdot)$.

[23] Recall that, in our empirical application, we assume each driver has a single possible route for their trip.

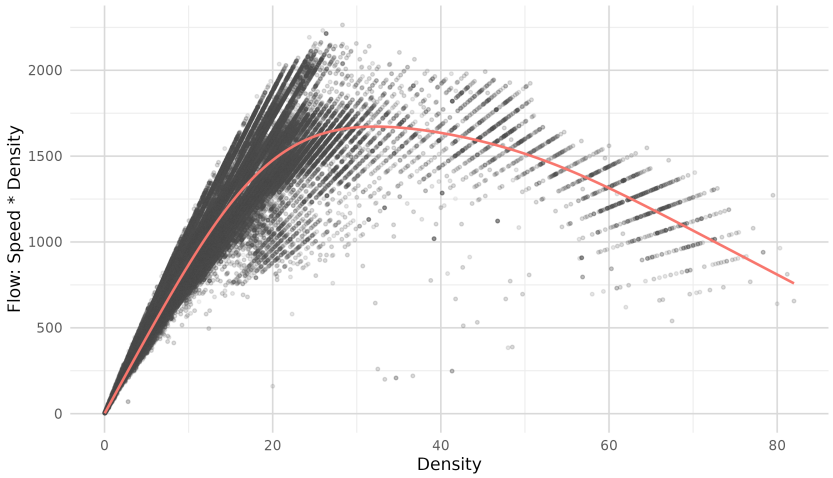


Figure 10: An estimated congestion function relating traffic flow on a segment of the Ayalon highway to traffic density on that segment.

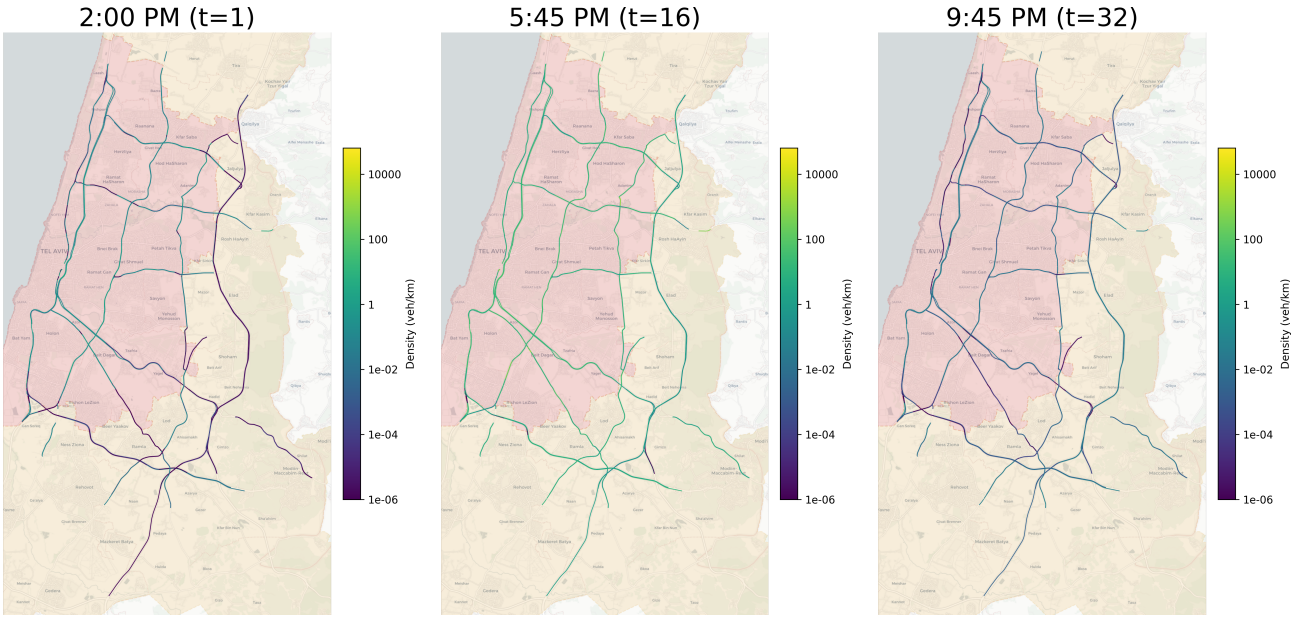


Figure 11: Equilibrium traffic between 2 PM and 10 PM without prices in the Central District

Baseline Equilibrium Figure 11 demonstrates the equilibrium flows at three points of afternoon/evening travel in the baseline when no congestion prices are charged. At 2PM traffic within the metro area (in shaded red) is moderate whereas traffic in the metro fringe (in shaded yellow) is low. During peak hours at 5:45, traffic throughout the Central District is high. At the end of the day at 9:45, traffic is relatively low and dispersed throughout the road network.

Experiment Fee Schedule We first consider how the fee schedule used in our experiment would affect equilibrium traffic and welfare if it were applied to all drivers in the Central District. Table 4 summarizes the difference in each driver's probability of driving (relative to the outside option), the average trip duration and price conditional on driving, and aggregate money-metric welfare with and

Metric	Mean	Mean %	SD	Q25	Median	Q75
Avg Trip Probability (%)	-0.03	-9.76%	0.07	-0.03	-0.01	0.00
Avg Trip Duration (mins)	-1.65	-9.38%	2.68	-2.47	-0.28	0.00
Avg Trip Price (NIS)	5.38		7.15	1.09	2.72	6.48
Unrebated Welfare (NIS)	-1.46	-10.93%	4.57	-1.26	-0.22	-0.02
Rebated Welfare (NIS)	0.41	3.07%	4.57	0.61	1.65	1.85

Table 4: Equilibrium effects of the experiment fee schedule on driver-level trip outcomes and money-metric welfare.

(a) Usage-based

	Peak	Off-peak
Metropolis	1.185	0.574
Metro. fringe	0.448	0.287
Periphery	0.000	0.000

(b) Cordon-based

	Peak	Off-peak
Metro. fringe → Metropolis	9.669	3.775
Metro. fringe → Periphery	0.453	0.080
Periphery → Metro. fringe	0.048	0.006

Table 5: Optimal price levels with full revenue recycling ($\Lambda = 1$) under usage-based and cordon-based pricing.

without an average toll rebate relative to the equilibrium without pricing. On average, the experiment fee schedule would reduce the probability of driving by 9.76%. For trips that are still taken, average duration would decrease by 9.38% at an average cost of 5.38 NIS. By comparison, New York City's congestion pricing program reportedly reduced traffic volumes by 11% (Badger et al., 2026) and increased average trip speeds on affected roads by 5-8% (Cook et al., 2025). All in all if tolls were fully rebated, then the experiment fee schedule would increase welfare by 3.07%. However, the rebates are necessary: if toll revenues were witheld from drivers, then average welfare would decrease by 10.93%.

Optimal Usage-Based Pricing We now consider how the equilibrium effects might differ if the prices in the experiment were set so as to maximize money-metric welfare. For tractability, we retain the pricing buckets used in the experiment—three categories of time and three categories of geography—and solve for welfare-maximizing prices for each bucket using the price optimization algorithm described in Appendix B.4. The left panel of Table 5 reports the resulting optimal usage-based price schedule when toll revenues are fully recycled. Compared to the experiment's fee schedule, the optimal price for kilometers during peak hours in metro areas are 21% lower, but the price during off-peak hours and in the metropolitan fringe would be several times higher. This reflects the idea that many of the congested highway segments in the central district are outside the metro region. Because the optimal prices target more of these segments, they remove substantially more trips altogether: the average probability of driving decreases by 26.38% relative to the equilibrium without pricing. However, this does not increase rebated welfare by all that much: average welfare increases by 1.15 percentage points from roughly 3% to 4% relative to the equilibrium without pricing. We do not report the optimal prices when toll revenues are discarded ($\Lambda = 0$) because they are all zeros. This suggests that the sorting gains from usage-based prices under these pricing buckets alone are not sufficient to offset the toll costs on their own. In an upcoming version of this draft, we compute the threshold value of Λ under which optimal usage-based prices are net welfare improving.

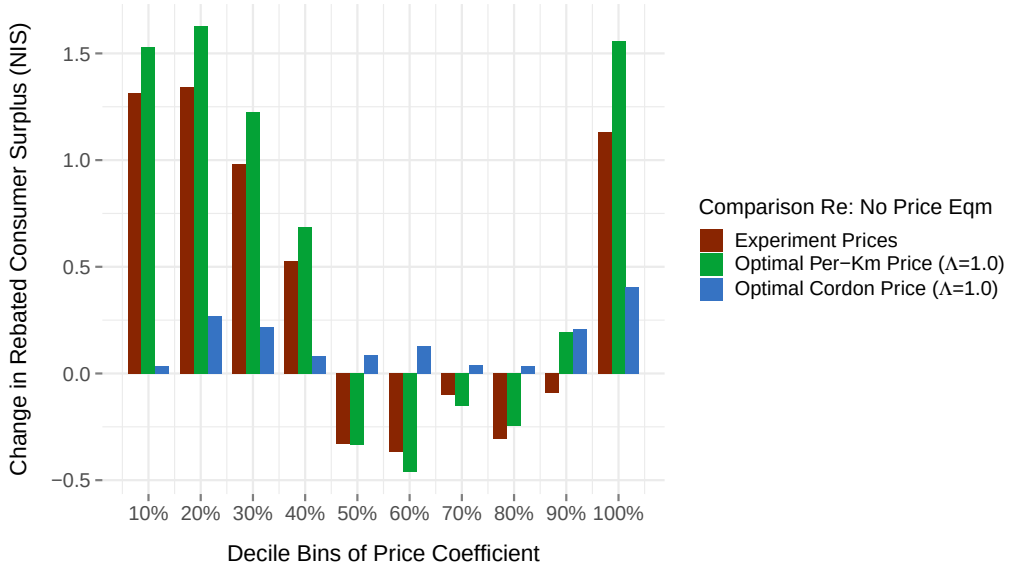


Figure 12: Distribution of per-driver rebated welfare by price coefficient decile under different pricing schemes.

Optimal Cordon Pricing In practice, most congestion pricing policies are implemented through cordon zones, rather than usage-based fees. As such, we also consider how optimal cordon pricing might differ from its usage-based analogue with the same objective, again using the price optimization algorithm described in Appendix B.4. We retain the same three categories of time and geography as in the usage-based pricing case, but model prices as binding only on segments that are at the entrance of each geographic zone. The right panel of Table 5 reports the corresponding optimal cordon schedule for directional zone-crossings with non-zero optimal tolls. Unsurprisingly, the largest tolls are at the crossing from the metropolitan fringe into the metro zone during peak hours: these are the trips that contribute most to congestion and that get priced highest under the usage based scheme. However, because Cordon pricing is less personalized, it induces smaller welfare changes overall. The average probability of driving decreases by only 5% relative to the equilibrium without pricing, and average welfare increases by 1.13%.

Distributional Effects Figure 12 shows the distribution of the difference in rebated per-driver welfare by price coefficient decile under the experiment's pricing, optimal usage-based and optimal cordon-based pricing. The two usage-based pricing schemes generally align directionally: drivers with the highest and lowest price coefficients experience the largest welfare gains, while drivers with intermediate price coefficients experience welfare losses. This makes intuitive sense: drivers with the lowest price coefficients benefit the most from the speed improvements, while drivers with the highest price coefficients benefit the most from the toll rebates. The cordon-based scheme, on the other hand, benefits every bin of drivers but generally induces smaller welfare changes overall.

References

Almagro, M., F. Barbieri, J. C. Castillo, N. G. Hickok, and T. Salz (2024): "Optimal Urban Transportation Policy: Evidence from Chicago," Tech. rep., National Bureau of Economic Research.

Arnott, R., A. De Palma, and R. Lindsey (1993): "A structural model of peak-period congestion: A traffic bottleneck with elastic demand," *The American Economic Review*, 161–179.

Badger, E., L. Buchanan, S. Chen, A. Elkeurti, and W. Hu (2026): "27 Million Fewer Car Trips: Life After a Year of Congestion Pricing," *The New York Times*.

Barth, M. and K. Boriboonsomsin (2008): "Real-world carbon dioxide impacts of traffic congestion," *Transportation research record*, 2058, 163–171.

Barwick, P. J., S. Li, A. Waxman, J. Wu, and T. Xia (2024): "Efficiency and equity impacts of urban transportation policies with equilibrium sorting," *American Economic Review*, 114, 3161–3205.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. (1994): "Estimating discrete-choice models of product differentiation," *The Rand journal of economics*, 242–262.

Borusyak, K., X. Jaravel, and J. Spiess (2021): "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*.

Boyles, S. D., N. E. Lownes, and A. Unnikrishnan (2025): "Transportation network analysis," *Volume I: Static and Dynamic Traffic Assignment*.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995): "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, 16, 1190–1208.

Callaway, B. and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of econometrics*, 225, 200–230.

Chandler-Wilde, H. (2023): "These Are the World's Most Congested Cities," *Bloomberg*.

Chernozhukov, V., I. Fernández-Val, and Y. Luo (2018): "The sorted effects method: discovering heterogeneous effects beyond their averages," *Econometrica*, 86, 1911–1938.

Cook, C., A. Kreidieh, S. Vasserman, H. Allcott, N. Arora, F. van Sambeek, A. Tomkins, and E. Turkel (2025): "The Short-Run Effects of Congestion Pricing in New York City," Working Paper 33584, National Bureau of Economic Research.

Cook, C. and P. Z. Li (2025): "Value Pricing or Lexus Lanes? The Distributional Effects of Dynamic Tolling," Tech. rep.

Currie, J. and R. Walker (2011): "Traffic congestion and infant health: Evidence from E-ZPass," *American Economic Journal: Applied Economics*, 3, 65–90.

Daganzo, C. F. (1994): "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation research part B: methodological*, 28, 269–287.

——— (1995): "The cell transmission model, part II: network traffic," *Transportation Research Part B: Methodological*, 29, 79–93.

Durrmeyer, I. and N. Martinez (2022): "The welfare consequences of urban traffic regulations," .

ERI SalaryExpert (2026): "Average Salary in Israel," .

Grieco, P. L. E., C. Murry, J. Pinkse, and S. Sagl (2025): "Optimal estimation of discrete choice demand models with consumer and product data," .

Hahn, R. W., R. D. Metcalfe, and E. Tam (2023): "Welfare Estimates of Shifting Peak Travel," Tech. rep., National Bureau of Economic Research.

Kahneman, D., A. B. Krueger, D. Schkade, N. Schwarz, and A. A. Stone (2006): "Would you be happier if you were richer?" *science*, 312, 1908–1910.

Kelley, C. T. (2022): *Solving Nonlinear Equations with Iterative Methods: Solvers and Examples in Julia*, vol. 20 of *Fundamentals of Algorithms*, SIAM.

Kreindler, G. (2024): "Peak-Hour Road Congestion Pricing: Experimental Evidence and Equilibrium Implications," *Econometrica*, 92, 1233–1268.

Lebacque, J. (1996): "The godunov scheme and what it means for rst order tra c ow models," in *Proceedings of the 13th International Symposium on Transportation and Trafc Theory, Lyon, France, July*, vol. 2426.

Martin, L. A. and S. Thornton (2017): "Can Road Charges Alleviate Congestion?" *Available at SSRN 3055428*.

Pigou, A. C. (1920): *The Economics of Welfare*, Macmillan and Co.

Revelt, D. and K. Train (2000): "Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier," .

Romano, J. P., A. M. Shaikh, and M. Wolf (2010): "Hypothesis testing in econometrics," *Annu. Rev. Econ.*, 2, 75–104.

Romano, J. P. and M. Wolf (2005): "Stepwise multiple testing as formalized data snooping," *Econometrica*, 73, 1237–1282.

Shao, J. and D. Tu (2012): *The jackknife and bootstrap*, Springer Science & Business Media.

Small, K. A. and X. Chu (2003): "Hypercongestion," *Journal of Transport Economics and Policy (JTEP)*, 37, 319–352.

Stewart, W. J. (1994): *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press.

Su, Y. (2022): "Measuring the value of urban consumption amenities: A time-use approach," *Journal of Urban Economics*, 132, 103495.

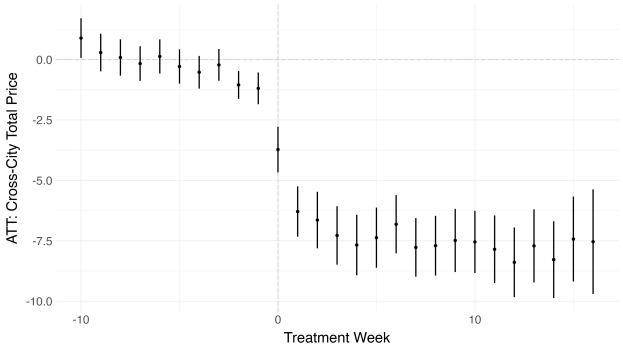
The Times of Israel (2021): "Worst-ever Passover traffic chaos as reopened Israel enjoys vaccinated vacation," *The Times of Israel*.

Train, K. E. (2009): *Discrete choice methods with simulation*, Cambridge university press.

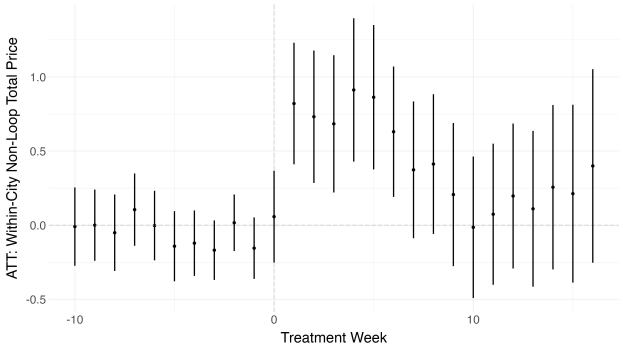
Vickrey, W. S. (1969): "Congestion theory and transport investment," *The American economic review*, 59, 251–260.

A Additional Tables and Figures

A.1 Event Studies

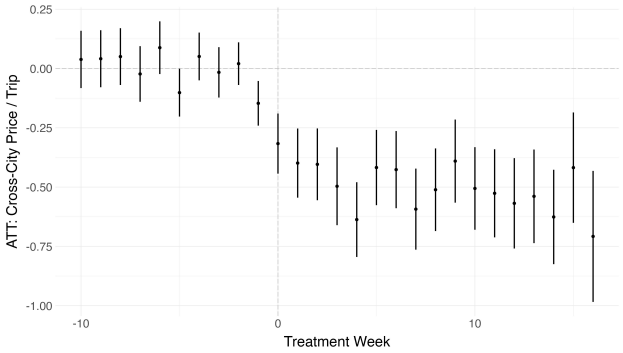


(a) Outcome: total price of cross-city trips per week

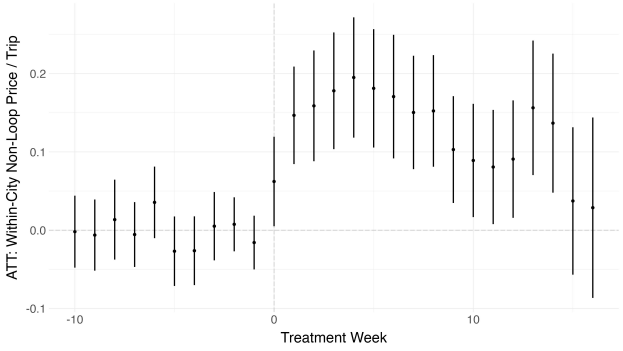


(b) Outcome: total price of within-city non-loop trips per week

Figure 13: Event studies of dynamic average treatment effects across weeks relative to treatment week.

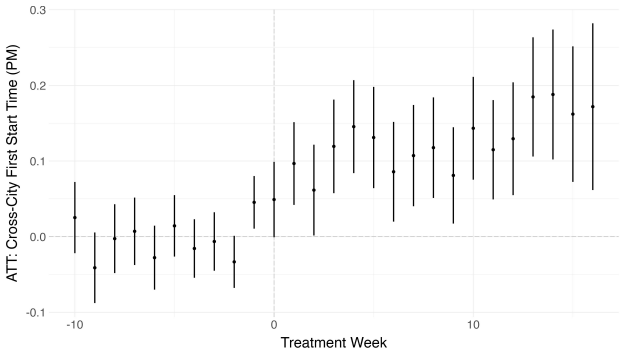


(a) Outcome: price per trip of cross-city trips per week

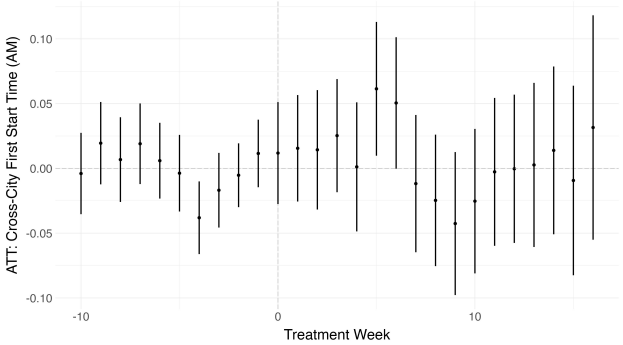


(b) Outcome: price per trip of within-city non-loop trips per week

Figure 14: Event studies of dynamic average treatment effects across weeks relative to treatment week.



(a) Outcome: first PM trip start time for cross-city trips



(b) Outcome: first AM trip start time for cross-city trips

Figure 15: Event studies of dynamic average treatment effects on departure times for cross-city trips.

B Equilibrium Computation and Price Optimization

In this appendix, we provide more details about how we discretize the model described in Section 7 to facilitate efficient equilibrium computation and congestion price optimization.

B.1 Model Discretization

First, we explain in detail how we discretize the components of the continuous-time model described in Section 7.

Period grid and frozen rates. We partition the horizon into M periods of common length Δ,

$$I_t \coloneqq [t\Delta, (t+1)\Delta), \qquad t = 0, \dots, M-1, \qquad T = M\Delta,$$

and represent each period by its average movement-level masses. For each movement (e, e'), define

$$\bar{q}_{ee't} \coloneqq \Delta^{-1} \int_{I_t} q_{ee'}(\tau)\, d\tau, \qquad \bar{q}_{et} \coloneqq \sum_{e' \in \mathcal{E} \cup \{\varnothing\}} \bar{q}_{ee't},$$

and collect the movement masses into the vector $\bar{\boldsymbol{q}}_t$. We then freeze the congestion-dependent transition rates within period t at these representative masses:

$$\lambda_{ee't} \coloneqq \lambda_{ee'}(\bar{\boldsymbol{q}}_t), \qquad Q_{jt} \coloneqq Q_j(\bar{\boldsymbol{q}}_t). \tag{9}$$

We likewise hold the inflow to route j constant within period t at rate a_{jt} and write $\boldsymbol{q}_{jt} \coloneqq \boldsymbol{q}_j(t\Delta)$ for the start-of-period route mass. The route-level ODE (6) then becomes, on I_t,

$$\frac{d}{d\tau}\boldsymbol{q}_j(\tau) = Q_{jt}\boldsymbol{q}_j(\tau) + a_{jt}\mathbf{e}_1. \tag{10}$$

Forward mass propagation. Because the coefficients are constant within a period, the route-level mass path has a closed-form solution. For local time $u \in [0, \Delta]$,

$$\boldsymbol{q}_j(t\Delta + u) = e^{Q_{jt}u}\boldsymbol{q}_{jt} + a_{jt}\left(\int_0^u e^{Q_{jt}s}\, ds\right)\mathbf{e}_1. \tag{11}$$

It is therefore convenient to define the block operators

$$E_{jt} \coloneqq e^{Q_{jt}\Delta}, \qquad A_{jt} \coloneqq \int_0^\Delta e^{Q_{jt}s}\, ds, \qquad B_{jt} \coloneqq \int_0^\Delta (\Delta - s)e^{Q_{jt}s}\, ds. \tag{12}$$

Evaluating (11) at $u = \Delta$ gives the start-of-next-period recursion,

$$\boldsymbol{q}_{j(t+1)} = E_{jt}\boldsymbol{q}_{jt} + a_{jt}A_{jt}\mathbf{e}_1, \tag{13}$$

while averaging (11) over the period gives the route-level period-average masses

$$\bar{\boldsymbol{q}}_{jt} \coloneqq \Delta^{-1} \int_{I_t} \boldsymbol{q}_j(\tau)\, d\tau = \Delta^{-1} \left[A_{jt} \boldsymbol{q}_{jt} + a_{jt} B_{jt} \mathbf{e}_1 \right]. \tag{14}$$

Backward value recursion. The backward pass mirrors the forward discretization. Using the generic notation from (7), let $\boldsymbol{X}_{jt}(\bar{\boldsymbol{q}})$ stack the start-of-period continuation values $X_{jkt}(\bar{\boldsymbol{q}})$ across states $k \in \{1, \ldots, K_j + 1\}$, with the absorbing component equal to zero. Suppose that within period t, being in segment-state k accrues running amount r_{jkt} per unit time, that a jump from state k to $k+1$ incurs lump-sum amount m_{jkt}, and that the post-horizon continuation values are collected in the vector $\bar{\boldsymbol{X}}_j$. Define the period-t source vector

$$\boldsymbol{c}^X_{jt} \coloneqq \left(r_{j1t} + Q_{jt}[2,1] m_{j1t},\, \ldots,\, r_{jK_jt} + Q_{jt}[K_j+1, K_j] m_{jK_jt},\, 0 \right)'.$$

For local time $u \in [0, \Delta]$, write $\boldsymbol{X}_j(t\Delta + u; \bar{\boldsymbol{q}})$ for the within-period continuation profile, so that $\boldsymbol{X}_j(t\Delta; \bar{\boldsymbol{q}}) = \boldsymbol{X}_{jt}(\bar{\boldsymbol{q}})$ and $\boldsymbol{X}_j((t+1)\Delta; \bar{\boldsymbol{q}}) = \boldsymbol{X}_{j(t+1)}(\bar{\boldsymbol{q}})$. Freezing the transition rates within the period turns (7) into

$$\frac{d}{du} \boldsymbol{X}_j(t\Delta + u; \bar{\boldsymbol{q}}) = -\boldsymbol{c}^X_{jt} - Q'_{jt} \boldsymbol{X}_j(t\Delta + u; \bar{\boldsymbol{q}}).$$

Left-multiplying by the integrating factor $e^{Q'_{jt}u}$ gives

$$\frac{d}{du} \left[e^{Q'_{jt}u} \boldsymbol{X}_j(t\Delta + u; \bar{\boldsymbol{q}}) \right] = -e^{Q'_{jt}u} \boldsymbol{c}^X_{jt}.$$

Integrating from 0 to Δ and rearranging yields the start-of-period backward recursion

$$\boldsymbol{X}_{jt}(\bar{\boldsymbol{q}}) = A'_{jt} \boldsymbol{c}^X_{jt} + E'_{jt} \boldsymbol{X}_{j(t+1)}(\bar{\boldsymbol{q}}). \tag{15}$$

This recursion is initialized at the horizon by the terminal condition $\boldsymbol{X}_{jM}(\bar{\boldsymbol{q}}) = \bar{\boldsymbol{X}}_j$. Expected duration and price are the two special cases used in the main text. For duration, $r_{jkt} = 1$, $m_{jkt} = 0$, and $\bar{\boldsymbol{X}}_j = \boldsymbol{D}^{\text{ff}}_{jM}$, so

$$\boldsymbol{c}^D_j \coloneqq (1, \ldots, 1, 0)'.$$

For price, $r_{jkt} = 0$, $m_{jkt} = p_{e_{jk}t}$, where $p_{et} \coloneqq p_e(t\Delta)$, and $\bar{\boldsymbol{X}}_j = \mathbf{0}$, which implies

$$\boldsymbol{c}^P_{jt} \coloneqq \left(p_{e_{j1}t} Q_{jt}[2,1],\, \ldots,\, p_{e_{jK_j}t} Q_{jt}[K_j+1, K_j],\, 0 \right)'.$$

Specializing (15) therefore gives

$$\boldsymbol{D}_{jt}(\bar{\boldsymbol{q}}) = A'_{jt} \boldsymbol{c}^D_j + E'_{jt} \boldsymbol{D}_{j(t+1)}(\bar{\boldsymbol{q}}), \qquad \boldsymbol{P}_{jt}(\bar{\boldsymbol{q}}) = A'_{jt} \boldsymbol{c}^P_{jt} + E'_{jt} \boldsymbol{P}_{j(t+1)}(\bar{\boldsymbol{q}}).$$

The route-period expected duration and price that enter demand are the first components of these vectors: $D_{jt}(\bar{\boldsymbol{q}}) = D_{j1t}(\bar{\boldsymbol{q}})$ and $P_{jt}(\bar{\boldsymbol{q}}) = P_{j1t}(\bar{\boldsymbol{q}})$.

Trip demand. Since scheduling preferences $\tilde{V}_{jt}(\theta)$ are assumed constant within each period, we let $\mathcal{T}_j(\theta) \subseteq [M]$ denote the set of feasible departure periods on route j. A route-period alternative (j,t) represents all departures in the interval I_t, so under the within-period constant-utility approximation, its continuous-time attractiveness is proportional to

$$\int_{I_t} \exp(\delta_j(\tau; \bar{\boldsymbol{q}}, \theta))\, d\tau \approx \Delta \exp\Big(\tilde{V}_{jt}(\theta) - \gamma D_{jt}(\bar{\boldsymbol{q}}) - \alpha P_{jt}(\bar{\boldsymbol{q}})\Big).$$

Given the assumption of unit-scale Type-I extreme-value shocks, this bin-width factor can be absorbed into the deterministic utility by redefining the scheduling preference term as $V_{jt}(\theta) \coloneqq \tilde{V}_{jt}(\theta) + \log \Delta$, i.e. just a common shift to inside alternatives relative to the outside option.

The discretized route-period utility index can then be written as

$$\delta_{jt}(\bar{\boldsymbol{q}}, \theta) \coloneqq V_{jt}(\theta) - \gamma D_{jt}(\bar{\boldsymbol{q}}) - \alpha P_{jt}(\bar{\boldsymbol{q}}),$$

which implies the type-θ probability of choosing route-period pair (j,t) is given by the typical logit choice probability formula:

$$s_{jt}(\bar{\boldsymbol{q}}, \theta) \coloneqq \frac{\exp(\delta_{jt}(\bar{\boldsymbol{q}}, \theta))}{1 + \sum_{j' \in \mathcal{J}(\theta)} \sum_{t' \in \mathcal{T}_{j'}(\theta)} \exp(\delta_{j't'}(\bar{\boldsymbol{q}}, \theta))}, \tag{16}$$

and that the aggregate share is given by

$$s_{jt}(\bar{\boldsymbol{q}}) \coloneqq \int_{\Theta} s_{jt}(\bar{\boldsymbol{q}}, \theta)\, dF(\theta).$$

Conditional on choosing (j,t), a driver's departure time is uniform on I_t, so the within-period inflow rate is

$$a_{jt}(\bar{\boldsymbol{q}}) \coloneqq \frac{Q\, s_{jt}(\bar{\boldsymbol{q}})}{\Delta}. \tag{17}$$

Discretized equilibrium. The discretized equilibrium is therefore a sequence $\{\bar{\boldsymbol{q}}_t^*\}_{t=0}^{M-1}$ such that, after solving the backward recursions, computing route-period demand, and propagating route-level masses forward, aggregating route-level period averages reproduces the original movement-level average masses:

$$\bar{q}^*_{ee't} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e,\, e_{j(k+1)} = e'\}\, \bar{q}^*_{jkt}, \qquad \forall\, (e,e'),\, t, \tag{18}$$

where we adopt the convention $e_{j(K_j+1)} \equiv \varnothing$ for the terminal movement.

B.2 Uniformization and Efficient Operator Evaluation

The recursions (13) and (15) require repeated applications of the operators E_{jt}, A_{jt}, and B_{jt} to a small set of vectors, along with the adjoint operators E'_{jt} and A'_{jt} in the backward pass. Directly evaluating matrix exponentials in every route-period block is unnecessary because each Q_{jt} is a CTMC generator with bidiagonal structure inherited from the route progression process.

Uniformization converts the matrix exponential into a Poisson-weighted power series (see Stewart (1994) for a textbook treatment). For each route-period pair (j,t), choose

$$\bar{\lambda}_{jt} \geq \max_{k\in[K_j]} \bigl(-Q_{jt}[k,k]\bigr), \qquad \Phi_{jt} \coloneqq I + \bar{\lambda}_{jt}^{-1} Q_{jt}, \qquad \mu_{jt} \coloneqq \bar{\lambda}_{jt}\Delta. \tag{19}$$

Then Φ_{jt} is a column-substochastic matrix, and for any $s \geq 0$,

$$e^{Q_{jt}s} = \sum_{n=0}^{\infty} w_n(\bar{\lambda}_{jt}s)\Phi_{jt}^n, \qquad w_n(\mu) \coloneqq e^{-\mu}\mu^n/n!. \tag{20}$$

Substituting (20) into the integral operators and using the fact that Φ_{jt} does not depend on s yields

$$a_n(\mu) \coloneqq \int_0^{\Delta} \frac{e^{-\bar{\lambda}_{jt}s}(\bar{\lambda}_{jt}s)^n}{n!}\, ds = \bar{\lambda}_{jt}^{-1} \Pr\bigl(\text{Poisson}(\mu) \geq n+1\bigr),$$

and

$$b_n(\mu) \coloneqq \int_0^{\Delta} (\Delta - s)\frac{e^{-\bar{\lambda}_{jt}s}(\bar{\lambda}_{jt}s)^n}{n!}\, ds = \Delta a_n(\mu) - \frac{n+1}{\bar{\lambda}_{jt}^2} \Pr\bigl(\text{Poisson}(\mu) \geq n+2\bigr).$$

As a result, the operator actions can be written as

$$E_{jt}v = \sum_{n=0}^{\infty} w_n(\mu_{jt})\Phi_{jt}^n v, \quad A_{jt}v = \sum_{n=0}^{\infty} a_n(\mu_{jt})\Phi_{jt}^n v, \quad B_{jt}v = \sum_{n=0}^{\infty} b_n(\mu_{jt})\Phi_{jt}^n v.$$

The backward pass uses the adjoint actions

$$E'_{jt}v = \sum_{n=0}^{\infty} w_n(\mu_{jt})(\Phi'_{jt})^n v, \qquad A'_{jt}v = \sum_{n=0}^{\infty} a_n(\mu_{jt})(\Phi'_{jt})^n v.$$

These series can all be evaluated in a single loop. Initialize $v^{(0)} = v$, $w_0 = e^{-\mu_{jt}}$, and $F_{-1} = 0$. For $n = 0, 1, 2, \ldots$, update either

$$v^{(n+1)} = \Phi_{jt}v^{(n)} \qquad \text{or} \qquad v^{(n+1)} = \Phi'_{jt}v^{(n)},$$

depending on whether the forward or adjoint action is needed, and update the scalar weights by

$$F_n = F_{n-1} + w_n, \quad w_{n+1} = w_n\mu_{jt}/(n+1), \quad a_n = \bar{\lambda}_{jt}^{-1}(1 - F_n), \quad b_n = \Delta a_n - \frac{n+1}{\bar{\lambda}_{jt}^2}(1 - F_n - w_{n+1}). \tag{21}$$

Accumulating the iterates with weights (w_n, a_n, b_n) gives the desired operator actions, and the loop can be truncated at the smallest $N_{jt}(\varepsilon)$ such that $1 - F_n \leq \varepsilon$. Because Φ_{jt} is bidiagonal, each multiplication $\Phi_{jt}v^{(n)}$ or $\Phi'_{jt}v^{(n)}$ costs only $\mathcal{O}(K_j)$ operations, so the total work per route in one outer fixed-point iteration is

$$\mathcal{O}\left(\sum_{t=0}^{M-1} K_j N_{jt}(\varepsilon)\right). \tag{22}$$

B.3 Equilibrium Computation

We solve the discretized fixed point (18) by iterating between route-level calculations and movement-level aggregation. In practice, we use simplex-constrained Anderson acceleration with damped updates to improve the stability of the outer fixed-point iteration and reduce the number of iterations needed for convergence by constructing better update directions from recent residuals (see Kelley (2022) for a textbook treatment). Let $\bar{\boldsymbol{q}}_t^{(r)}$ denote the vector of period-t average movement masses at outer iteration r.

1. **Initialize.** Choose an initial guess $\{\bar{\boldsymbol{q}}_t^{(0)}\}_{t=0}^{M-1}$, an Anderson memory length m, a damping parameter $\omega \in (0, 1]$, and a convergence tolerance $\zeta > 0$.

2. **Build generators.** For each route-period pair (j, t), construct the frozen generator from (9),

$$Q_{jt}^{(r)} \coloneqq Q_j(\bar{\boldsymbol{q}}_t^{(r)}),$$

   along with the associated uniformization objects $\bar{\lambda}_{jt}^{(r)}$, $\Phi_{jt}^{(r)}$, and $\mu_{jt}^{(r)}$ from (19).

3. **Backward pass.** Starting from $\boldsymbol{D}_{jM}^{(r)} = \boldsymbol{D}_{jM}^{\mathrm{ff}}$ and $\boldsymbol{P}_{jM}^{(r)} = \mathbf{0}$, iterate backward over $t = M - 1, \ldots, 0$ and each route j using the backward recursion (15) to compute

$$\boldsymbol{D}_{jt}^{(r)} = A_{jt}^{(r)\prime} \boldsymbol{c}_j^D + E_{jt}^{(r)\prime} \boldsymbol{D}_{j(t+1)}^{(r)}, \qquad \boldsymbol{P}_{jt}^{(r)} = A_{jt}^{(r)\prime} \boldsymbol{c}_{jt}^P + E_{jt}^{(r)\prime} \boldsymbol{P}_{j(t+1)}^{(r)}.$$

4. **Demand and inflows.** Using (16) and (17), compute route-period shares $s_{jt}^{(r)}$ and inflow rates

$$a_{jt}^{(r)} = Q\, s_{jt}^{(r)} / \Delta.$$

5. **Forward pass.** Set $\boldsymbol{q}_{j0}^{(r)} = \mathbf{0}$ for every route j. Then, for $t = 0, \ldots, M-1$, use the operators from (12), the forward recursion (13), and the period-average formula (14) to compute

$$\boldsymbol{q}_{j(t+1)}^{(r)} = E_{jt}^{(r)} \boldsymbol{q}_{jt}^{(r)} + a_{jt}^{(r)} A_{jt}^{(r)} \mathbf{e}_1,$$

   and

$$\bar{\boldsymbol{q}}_{jt}^{(r,\mathrm{new})} = \Delta^{-1} \left[A_{jt}^{(r)} \boldsymbol{q}_{jt}^{(r)} + a_{jt}^{(r)} B_{jt}^{(r)} \mathbf{e}_1 \right].$$

6. **Aggregate to movements.** Convert route-level period averages into movement-level averages using the aggregation map in (18):

$$\bar{q}_{ee't}^{(r,\mathrm{new})} \coloneqq \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e,\, e_{j(k+1)} = e'\}\, \bar{q}_{jkt}^{(r,\mathrm{new})}.$$

7. **Get update direction.** Stack the movement-level averages across periods and define the current raw fixed-point residual

$$\boldsymbol{g}^{(r)} \coloneqq \bar{\boldsymbol{q}}^{(r,\mathrm{new})} - \bar{\boldsymbol{q}}^{(r)}.$$

Store $\boldsymbol{g}^{(r)}$ in a rolling Anderson history. If the history now contains more than m entries, discard the oldest one so that only the most recent m residuals are retained. Let $\mathcal{H}_r$ denote the set of retained iteration indices. Simplex-constrained Anderson acceleration then chooses convex-combination weights

$$\boldsymbol{\alpha}^{(r)} \in \arg\min_{\boldsymbol{\alpha}\in\Delta^{|\mathcal{H}_r|}} \left\| \sum_{\ell\in\mathcal{H}_r} \alpha_\ell \boldsymbol{g}^{(\ell)} \right\|_2^2, \qquad \Delta^{|\mathcal{H}_r|} := \left\{ \boldsymbol{\alpha} \in \mathbb{R}_+^{|\mathcal{H}_r|} : \sum_{\ell\in\mathcal{H}_r} \alpha_\ell = 1 \right\},$$

and uses the resulting accelerated direction

$$\boldsymbol{d}^{(r)} := \sum_{\ell\in\mathcal{H}_r} \alpha_\ell^{(r)} \boldsymbol{g}^{(\ell)}.$$

Equivalently, period by period,

$$\boldsymbol{d}_t^{(r)} = \sum_{\ell\in\mathcal{H}_r} \alpha_\ell^{(r)} \left(\bar{\boldsymbol{q}}_t^{(\ell,\text{new})} - \bar{\boldsymbol{q}}_t^{(\ell)} \right).$$

When the history contains only the current residual, this reduces to the usual fixed-point residual.

8. **Damped update.** Update the movement masses according to

$$\bar{\boldsymbol{q}}_t^{(r+1)} := \bar{\boldsymbol{q}}_t^{(r)} + \omega \boldsymbol{d}_t^{(r)}, \qquad \forall\, t.$$

9. **Convergence test.** Stop when

$$\max_{\substack{t \\ e\to e'}} \left| \bar{q}_{ee't}^{(r,\text{new})} - \bar{q}_{ee't}^{(r)} \right| \le \zeta;$$

otherwise increment r and return to step 2.

At convergence, the movement-level average masses are consistent with the backward duration and price recursions, route-period demand, and forward propagation of route-level traffic through the network.

B.4 Price Optimization

We now describe how we choose a time-varying price schedule to maximize welfare while accounting for the endogenous equilibrium response of traffic.

Continuous-time welfare problem. Let $\boldsymbol{p} := \{p_e(\tau)\}_{e\in\mathcal{E},\,\tau\in[0,T]}$ denote a candidate collection of segment-time prices, and let $G(\boldsymbol{q},\boldsymbol{p})$ denote the continuous-time equilibrium map from Section 7, now written with its dependence on prices made explicit. For a given price path $\boldsymbol{p}$, the induced equilibrium movement masses $\boldsymbol{q}^*(\cdot;\boldsymbol{p})$ satisfy

$$G(\boldsymbol{q}^*(\cdot;\boldsymbol{p}),\boldsymbol{p}) = \boldsymbol{q}^*(\cdot;\boldsymbol{p}).$$

For a driver of type $\theta = (\mathcal{J}, \mathcal{T}, V(\cdot), \alpha, \gamma)$, define the systematic utility index

$$\delta_j(\tau; \boldsymbol{q}, \boldsymbol{p}, \theta) \coloneqq V_j(\tau; \theta) - \gamma D_j(\tau; \boldsymbol{q}) - \alpha P_j(\tau; \boldsymbol{q}, \boldsymbol{p}),$$

where we now write expected trip price as depending explicitly on both the equilibrium masses and the candidate price schedule. Let $s_j(\tau; \boldsymbol{q}, \boldsymbol{p})$ denote the aggregate route-departure-time share density from (5) after making the price dependence explicit, so that $a_j(\tau; \boldsymbol{q}, \boldsymbol{p}) = Q s_j(\tau; \boldsymbol{q}, \boldsymbol{p})$. Using the standard logit inclusive-value formula, type-θ consumer surplus under prices $\boldsymbol{p}(\cdot)$ is

$$CS(\theta; \boldsymbol{p}) \coloneqq \frac{1}{\alpha} \log \left(1 + \sum_{j \in \mathcal{J}(\theta)} \int_{\underline{T}_j(\theta)}^{\overline{T}_j(\theta)} \exp(\delta_j(u; \boldsymbol{q}^*(\cdot; \boldsymbol{p}), \boldsymbol{p}, \theta))\, du \right),$$

and aggregate expected toll payments are

$$R(\boldsymbol{p}) \coloneqq Q \sum_{j=1}^{J} \int_0^T s_j(\tau; \boldsymbol{q}^*(\cdot; \boldsymbol{p}), \boldsymbol{p}) P_j(\tau; \boldsymbol{q}^*(\cdot; \boldsymbol{p}), \boldsymbol{p})\, d\tau.$$

We therefore write the planner's continuous-time problem as

$$\begin{aligned} \max_{\boldsymbol{p}(\cdot) \in \mathcal{P}} \quad & W(\boldsymbol{p}) \coloneqq Q \int_\Theta CS(\theta; \boldsymbol{p})\, dF(\theta) + \Lambda R(\boldsymbol{p}) \\ \text{s.t.} \quad & G(\boldsymbol{q}^*(\cdot; \boldsymbol{p}), \boldsymbol{p}(\cdot)) = \boldsymbol{q}^*(\cdot; \boldsymbol{p}), \end{aligned} \tag{23}$$

where $\mathcal{P}$ is the admissible set of price paths, for example one imposing pointwise lower and upper bounds on each segment-time price, and $\Lambda \geq 0$ governs the planner's weight on revenue relative to consumer surplus.

Discretized price optimization problem. Under the discretization described above, prices are summarized by the segment-period values $\bar{\boldsymbol{p}} \coloneqq \{p_{et}\}_{e,t}$. Given candidate prices $\bar{\boldsymbol{p}}$, the backward recursion for expected prices from (15) becomes

$$\boldsymbol{P}_{jt}(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}}) = A'_{jt} \boldsymbol{c}^P_{jt}(\bar{\boldsymbol{p}}) + E'_{jt} \boldsymbol{P}_{j(t+1)}(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}}),$$

where

$$\boldsymbol{c}^P_{jt}(\bar{\boldsymbol{p}}) \coloneqq \left(p_{e_{j1}t} Q_{jt}[2, 1],\, \ldots,\, p_{e_{jK_j}t} Q_{jt}[K_j + 1, K_j],\, 0 \right)'.$$

Let $G(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}})$ now denote the discretized movement-level map that takes candidate movement masses and prices into the updated movement masses obtained by the backward pass, the route-period demand system, the forward pass, and movement-level aggregation described in Appendices B.1 and B.3. Then the discretized equilibrium associated with $\bar{\boldsymbol{p}}$ satisfies

$$G(\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}), \bar{\boldsymbol{p}}) = \bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}).$$

For type θ, the discretized utility index is

$$\delta_{jt}(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}}, \theta) \coloneqq V_{jt}(\theta) - \gamma D_{jt}(\bar{\boldsymbol{q}}) - \alpha P_{jt}(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}}),$$

and let $s_{jt}(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}})$ denote the aggregate route-period shares from (16) after making the dependence on candidate prices explicit. The discretized type-θ consumer surplus and aggregate expected payment are therefore

$$CS(\theta; \bar{\boldsymbol{p}}) = \frac{1}{\alpha} \log\left(1 + \sum_{j \in \mathcal{J}(\theta)} \sum_{t \in \mathcal{T}_j(\theta)} \exp(\delta_{jt}(\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}), \bar{\boldsymbol{p}}, \theta))\right),$$
$$R(\bar{\boldsymbol{p}}) = Q \sum_{j=1}^{J} \sum_{t=0}^{M-1} s_{jt}(\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}), \bar{\boldsymbol{p}}) P_{jt}(\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}), \bar{\boldsymbol{p}}).$$

In the implementation, prices may be parameterized by a lower-dimensional vector $\boldsymbol{\vartheta}$ that maps linearly into segment-period prices $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})$; optimizing every segment-period price separately is the special case $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}) = \boldsymbol{\vartheta}$. The discretized problem is therefore

$$\begin{aligned} \max_{\underline{\boldsymbol{\vartheta}} \le \boldsymbol{\vartheta} \le \overline{\boldsymbol{\vartheta}}} \quad & \bar{W}(\boldsymbol{\vartheta}) \coloneqq \int_{\Theta} CS(\theta; \bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))\, dF(\theta) + \Lambda R(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) \\ \text{s.t.} \quad & G(\bar{\boldsymbol{q}}^*(\boldsymbol{\vartheta}), \bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) = \bar{\boldsymbol{q}}^*(\boldsymbol{\vartheta}). \end{aligned} \tag{24}$$

Optimization algorithm. We solve the optimization problem (24) using the box-constrained first-order optimization algorithm L-BFGS-B (Byrd et al., 1995), treating the equilibrium masses as an implicit function of the prices. At any candidate parameter vector $\boldsymbol{\vartheta}$, define $\bar{\boldsymbol{p}} = \bar{\boldsymbol{p}}(\boldsymbol{\vartheta})$ and let $\bar{\boldsymbol{q}}^*(\boldsymbol{\vartheta})$ denote the solution to the discretized equilibrium fixed point. The total welfare gradient therefore takes the form

$$\nabla_{\boldsymbol{\vartheta}} \bar{W} = \partial_{\boldsymbol{\vartheta}} \bar{W} + \partial_{\bar{\boldsymbol{q}}} \bar{W} \frac{d\bar{\boldsymbol{q}}^*}{d\boldsymbol{\vartheta}},$$

where the first term is the direct effect of changing prices holding equilibrium masses fixed, and the second term captures how welfare changes because the equilibrium itself moves with prices. To characterize $d\bar{\boldsymbol{q}}^*/d\boldsymbol{\vartheta}$, differentiate the fixed-point condition

$$G(\bar{\boldsymbol{q}}^*(\boldsymbol{\vartheta}), \bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) = \bar{\boldsymbol{q}}^*(\boldsymbol{\vartheta}),$$

which yields the implicit-function equation

$$[\partial_{\bar{\boldsymbol{q}}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}) - I] \frac{d\bar{\boldsymbol{q}}^*}{d\boldsymbol{\vartheta}} = -\partial_{\boldsymbol{\vartheta}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}(\boldsymbol{\vartheta})), \tag{25}$$

so a direct application would require solving for the equilibrium sensitivity of every price parameter separately. Since the equilibrium state vector $\bar{\boldsymbol{q}}^*$ is high dimensional, this is expensive. Instead, we use the adjoint form of the implicit function theorem, which introduces an adjoint vector $\boldsymbol{\lambda}$ satisfying

$$[\partial_{\bar{\boldsymbol{q}}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}) - I]' \boldsymbol{\lambda} = -\left(\partial_{\bar{\boldsymbol{q}}} \bar{W}\right)', \tag{26}$$

and then implies

$$\nabla_{\boldsymbol{\vartheta}} \bar{W} = \partial_{\boldsymbol{\vartheta}} \bar{W} + \left(\partial_{\boldsymbol{\vartheta}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))\right)' \boldsymbol{\lambda}.$$

This reformulation is useful here because it replaces one linear solve per price parameter with a single high-dimensional adjoint solve. Moreover, that adjoint system can be handled with an iterative linear solver using repeated applications of the linearized forward map and its adjoint, which is much easier than explicitly forming or factorizing the full Jacobian of the equilibrium system.

The resulting optimization routine proceeds as follows.

1. **Initialize.** Choose an initial price-parameter vector $\boldsymbol{\vartheta}^{(0)}$ satisfying the box constraints and an initial equilibrium guess. In practice, we warm-start the inner equilibrium solve with the converged equilibrium from the closest previously solved price vector, which for nearby L-BFGS-B iterates is typically just the previous optimizer iterate.

2. **Expand prices.** Map $\boldsymbol{\vartheta}^{(r)}$ into segment-period prices $\bar{\boldsymbol{p}}^{(r)} = \bar{\boldsymbol{p}}(\boldsymbol{\vartheta}^{(r)})$.

3. **Solve the inner equilibrium.** Holding $\bar{\boldsymbol{p}}^{(r)}$ fixed, solve

   $$G(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}}^{(r)}) = \bar{\boldsymbol{q}}$$

   using the equilibrium algorithm from Appendix B.3 to obtain $\bar{\boldsymbol{q}}^{*(r)}$.

4. **Evaluate welfare.** Using $\bar{\boldsymbol{q}}^{*(r)}$, compute the backward price recursion, route-period choice probabilities, and the welfare objective $\bar{W}(\boldsymbol{\vartheta}^{(r)})$ from (24).

5. **Evaluate the gradient.** Differentiate through the equilibrium map using (25). In practice, we solve the adjoint system (26) with an iterative linear solver like GMRES that only requires repeated evaluations of the linearized forward pass of the equilibrium system and its adjoint, and then assemble $\nabla_{\boldsymbol{\vartheta}} \bar{W}(\boldsymbol{\vartheta}^{(r)})$.

6. **Take an L-BFGS-B step.** Provide $-\bar{W}(\boldsymbol{\vartheta}^{(r)})$ and its gradient to box-constrained L-BFGS-B, which updates the price parameters subject to the lower and upper bounds to obtain $\boldsymbol{\vartheta}^{(r+1)}$.

7. **Convergence test.** Stop when the projected gradient norm or the relative change in the parameter vector or objective falls below the requested tolerance; otherwise increment r and return to step 2.

C Demand Estimation Details

In this appendix, we describe the details of the procedure we use to estimate trip demand in Section 8.1.

C.1 Estimator

We estimate the demand parameters by minimizing a negative panel log-likelihood with an IV-GMM penalty, akin to Grieco et al. (2025):

$$\begin{aligned} \hat{\theta} &\coloneqq \arg\min_{\gamma_0,\alpha_0,\Pi,L} \min_{\delta} Q(\delta,\gamma_0,\alpha_0,\Pi,L), \\ Q(\delta,\gamma_0,\alpha_0,\Pi,L) &\coloneqq -\frac{1}{N}\sum_{i=1}^{N} \log p_i(\delta,\gamma_0,\alpha_0,\Pi,L) + \frac{1}{N}\mathcal{P}(\delta,\gamma_0). \end{aligned} \tag{27}$$

Here, δ collects the market-level utility terms, γ_0 denotes the coefficient on market-level average duration that enters that utility, α_0 is the baseline price coefficient, Π collects the coefficients on demographic interactions, L parameterizes the covariance matrix of random coefficients, $p_i(\delta,\gamma_0,\alpha_0,\Pi,L)$ is driver i's panel marginal likelihood after integrating over unobserved taste heterogeneity, and $\mathcal{P}(\delta,\gamma_0)$ is an IV-GMM penalty induced by the market-level utility decomposition. The next paragraph defines the driver-specific likelihood contribution together with the market-level utility decomposition, and the paragraph after that shows how this decomposition generates the penalty.

Driver-specific marginal likelihood. It is convenient to write the driver-specific duration and price coefficients as

$$\begin{pmatrix} \gamma(X_i,\nu_i) \\ \alpha(X_i,\nu_i) \end{pmatrix} \coloneqq \begin{pmatrix} \gamma_0 \\ \alpha_0 \end{pmatrix} + \Pi X_i + L\nu_i, \quad \nu_i \sim \mathcal{N}(0, I_2),$$

where (γ_0,α_0) are baseline taste coefficients, the matrix Π collects the coefficients on characteristic interactions with demographics X_i, and $\Sigma = LL'$ is the covariance matrix of the random coefficients written in terms of its lower-triangular Cholesky factor L.

As discussed in Section 8, we only allow market-mean duration, and not market-mean price, to enter market-level utility. Writing $\overline{D}_{mt}$ for the average of $\hat{D}_{i\ell t}$ across trip opportunities in market-period (m,t), we decompose market-level utility as

$$\delta_{mt} = \hat{V}_{mt} + \gamma_0 \overline{D}_{mt}, \qquad \hat{V}_{mt} = \delta_{mt} - \gamma_0 \overline{D}_{mt}, \tag{28}$$

where $\hat{V}_{mt}$ is the component of market-level utility that captures scheduling preferences and other unobserved determinants of demand for departing in period t in market m.

Conditional on a draw of ν, we write the systematic utility of departure time t on date ℓ as

$$u_{i\ell t}(\delta,\gamma_0,\alpha_0,\Pi,L,\nu) \coloneqq \delta_{m(i,\ell)t} + \gamma_0\left(\hat{D}_{i\ell t} - \overline{D}_{m(i,\ell)t}\right) + \left(\gamma(X_i,\nu) - \gamma_0\right)\hat{D}_{i\ell t} + \alpha(X_i,\nu)\hat{P}_{i\ell t},$$

with the outside option normalized to $u_{i\ell\infty} = 0$. This separates the market-level component of utility, $\delta_{m(i,\ell)t}$, from the driver-specific component driven by within-market variation in duration and price and by heterogeneity in coefficients. By (28), this representation is exactly equivalent to writing the

utility index $u_{i\ell t}$ as

$$u_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu) = \hat{V}_{m(i,\ell)t} + \gamma(X_i, \nu)\hat{D}_{i\ell t} + \alpha(X_i, \nu)\hat{P}_{i\ell t}.$$

The corresponding conditional choice probabilities are

$$s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu) \coloneqq \frac{\exp(u_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu))}{1 + \sum_{t' \in \mathcal{T}_i} \exp(u_{i\ell t'}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu_i))},$$

with outside-option (indexed by $t = \infty$) probability,

$$s_{i\ell\infty}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu) \coloneqq 1 - \sum_{t \in \mathcal{T}_i} s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu).$$

Letting $Y_{i\ell\infty} \coloneqq 1 - \sum_{t \in \mathcal{T}_i} Y_{i\ell t}$ be an indicator of whether driver i chooses their outside option on date ℓ, driver i's marginal likelihood contribution is

$$p_i(\delta, \gamma_0, \alpha_0, \Pi, L) \coloneqq \int \prod_{\ell \in \mathcal{L}_i} \left[\prod_{t \in \mathcal{T}_i} s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu)^{Y_{i\ell t}} \cdot s_{i\ell\infty}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu)^{Y_{i\ell\infty}} \right] \phi(\nu)\, d\nu, \quad (29)$$

where ϕ is the standard bivariate normal density.

Market-level utility decomposition. If Z denotes the matrix of market-by-time-period level instruments described in Section 8, the decomposition in (28) implies the IV-GMM penalty

$$\mathcal{P}(\delta, \gamma_0) \coloneqq \tfrac{1}{2}\hat{V}' Z (Z'Z)^{-1} Z' \hat{V}.$$

Because price is excluded from (28), the mean price coefficient and the remaining components of the price-coefficient distribution are identified entirely from within-driver panel variation in prices.

C.2 Simulated Likelihood and Optimization

To optimize the objective function Q in (27) in practice, we use several techniques to improve computational efficiency and stability that we describe below.

Importance sampling for likelihood evaluation. The integral in (29) has no closed form, so we evaluate it by importance sampling. For a given driver, the integrand is the prior density over tastes multiplied by the likelihood of that driver's entire panel of choices. Because this likelihood multiplies together many date-level choice probabilities, values of ν that fit the observed sequence of choices even modestly better receive much more weight than values that fit them poorly. As a result, once a driver is observed making many repeated choices, the integrand becomes sharply peaked around the region of ν values that best rationalize that driver's behavior. Naive simulation from the diffuse prior is then inefficient because most draws fall in regions where the integrand is essentially zero. Instead,

we use importance sampling to concentrate ν_i draws in the region of ν values that matter most for the driver's likelihood contribution.

In particular, for a fixed value of $(\delta, \gamma_0, \alpha_0, \Pi, L)$, define driver i's conditional log-likelihood as

$$\ell_i(\nu) \coloneqq \sum_{\ell \in \mathcal{L}_i} \left[\sum_{t \in \mathcal{T}_i} Y_{i\ell t} \log s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu) + Y_{i\ell\infty} \log s_{i\ell\infty}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu) \right]. \tag{30}$$

The integrand in (29) is proportional to $\exp\big(\ell_i(\nu) - \frac{1}{2}\|\nu\|^2\big)$, which is exactly the posterior kernel for ν_i up to a normalizing constant (see Revelt and Train (2000) and Chapter 11 in Train (2009)). Hence, the part of the integral that matters most is the part of the posterior density with the most mass. A good proposal distribution for importance sampling should therefore mimic that posterior as closely as possible, while being cheap to evaluate since we do so in every iteration of the outer optimization loop.

We use a Laplace approximation to the posterior as the proposal distribution. Specifically, if $\hat{\nu}_i$ denotes the maximizer of $\ell_i(\nu) - \frac{1}{2}\|\nu\|^2$, then a second-order Taylor expansion around $\hat{\nu}_i$ gives

$$\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2 \approx \ell_i(\hat{\nu}_i) - \tfrac{1}{2}\|\hat{\nu}_i\|^2 - \tfrac{1}{2}(\nu - \hat{\nu}_i)' H_i (\nu - \hat{\nu}_i),$$

so the posterior is locally Gaussian with mean $\hat{\nu}_i$ and covariance matrix H_i^{-1} (see Chapter 11 in Train (2009)). Therefore, we use a Gaussian proposal distribution with mean $\hat{\nu}_i$ and covariance matrix H_i^{-1} as the importance-sampling proposal.

Key to our approach is that the posterior mode $\hat{\nu}_i$ and Hessian H_i can be computed very efficiently and in parallel across drivers i. To do so, we solve the optimization problem for the posterior mode using Newton method:

$$\hat{\nu}_i = \arg\max_{\nu} \left[\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2 \right]. \tag{31}$$

At each Newton step, the gradient and negative Hessian of the log posterior are

$$\nabla_\nu \left[\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2 \right] = L' \frac{\partial \ell_i(\nu)}{\partial \beta_i} - \nu, \quad H_i(\nu) = I_2 - L' \frac{\partial^2 \ell_i(\nu)}{\partial \beta_i \partial \beta_i'} L, \tag{32}$$

where the derivatives of $\ell_i(\nu)$ are taken with respect to $\beta_i = (\gamma_i, \alpha_i)'$, evaluated at $\beta_i = (\gamma_0, \alpha_0)' + \Pi X_i + L\nu$. Evaluated at the mode, we write

$$H_i \coloneqq H_i(\hat{\nu}_i) = U_i' U_i$$

for its Cholesky factorization.

Starting from common base draws $\{z_r\}_{r=1}^R$ from $\mathcal{N}(0, I_2)$, we then define the driver-specific proposal draws

$$\nu_{ir} \coloneqq \hat{\nu}_i + U_i^{-1} z_r, \qquad r = 1, \ldots, R. \tag{33}$$

Each draw receives log importance weight

$$\log w_{ir} \coloneqq \ell_i(\nu_{ir}) - \tfrac{1}{2}\|\nu_{ir}\|^2 + \log|\det U_i^{-1}| + \tfrac{1}{2}\|z_r\|^2. \tag{34}$$

The first two terms are the log of the unnormalized posterior density at the draw, while the final two terms divide out the Gaussian proposal density induced by (33). The corresponding normalized importance weights are

$$a_{ir} \coloneqq \frac{\exp(\log w_{ir})}{\sum_{r'=1}^{R} \exp(\log w_{ir'})}. \tag{35}$$

We then approximate driver i's log-likelihood contribution by

$$\log \hat{p}_i = \log\left(\sum_{r=1}^{R} e^{\log w_{ir}}\right) - \log R, \tag{36}$$

evaluated numerically using the standard log-sum-exp transformation for stability.

Computing gradients. Differentiating the simulated log-likelihood approximation in (36), which approximates the marginal likelihood in (29), and using the normalized importance weights in (35) yields the generic identity

$$\frac{\partial \log \hat{p}_i}{\partial \psi} = \sum_{r=1}^{R} a_{ir} \frac{\partial \ell_i(\nu_{ir})}{\partial \psi} \tag{37}$$

for any parameter ψ entering the utility index $u_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu)$. As such, we can write the gradient of the marginal likelihood as an Laplace-approximated posterior-weighted average of conditional logit scores.

For δ_{mt}, note first that, for any draw ν,

$$\frac{\partial \ell_i(\nu)}{\partial \delta_{mt}} = \sum_{\ell \in \mathcal{L}_i:\, m(i,\ell)=m} \left(Y_{i\ell t} - s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu)\right),$$

because in the logit model

$$\frac{\partial \log s_{i\ell t'}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu)}{\partial u_{i\ell t}} = 1\{t' = t\} - s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu).$$

Substituting this conditional score into the identity above gives

$$\frac{\partial \log \hat{p}_i}{\partial \delta_{mt}} = \sum_{\ell \in \mathcal{L}_i:\, m(i,\ell)=m} \sum_{r=1}^{R} a_{ir} \left(Y_{i\ell t} - s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu_{ir})\right).$$

The gradient of Q in (27) with respect to δ is therefore

$$\nabla_\delta Q = -\frac{1}{N} \sum_{i=1}^{N} \nabla_\delta \log \hat{p}_i + \frac{1}{N} P_Z \hat{V}. \tag{38}$$

The same argument gives the gradients of Q in (27) with respect to α_0 and Π. Define

$$z_{i\ell t} \coloneqq \left(\hat{P}_{i\ell t}, \hat{D}_{i\ell t} X_i', \hat{P}_{i\ell t} X_i'\right)'.$$

Then

$$\frac{\partial \log \hat{p}_i}{\partial(\alpha_0, \mathrm{vec}(\Pi))} = \sum_{\ell \in \mathcal{L}_i} \sum_{t \in \mathcal{T}_i} \sum_{r=1}^{R} a_{ir} \left(Y_{i\ell t} - s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu_{ir})\right) z_{i\ell t}. \tag{39}$$

Here the first component corresponds to α_0, while the remaining components correspond to the duration and price interaction rows of Π.

Finally, let $x_{i\ell t}^{\mathrm{rand}}$ denote the vector of characteristics with random coefficients; in our baseline specification these are just trip duration and price. Since $L\nu_{ir}$ shifts the random-coefficient vector in the direction of ν_{ir}, the same posterior-average-score calculation implies

$$\frac{\partial \log \hat{p}_i}{\partial L_{kq}} = \sum_{r=1}^{R} a_{ir} \sum_{\ell \in \mathcal{L}_i} \sum_{t \in \mathcal{T}_i} \left(Y_{i\ell t} - s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu_{ir})\right) x_{i\ell t, k}^{\mathrm{rand}} \nu_{ir, q}. \tag{40}$$

Because the IV-GMM penalty depends only on (δ, γ_0), the outer-loop gradient with respect to (α_0, Π, L) is simply the negative sample average of these likelihood gradients. We do not optimize over γ_0 by gradient descent; instead, it is updated by the closed-form IV regression step in the algorithm below.

Optimization algorithm. We solve the estimator by a nested algorithm that profiles out the high-dimensional mean-utility vector δ and updates γ_0 in closed form. Let $(\alpha_0^{(k)}, \Pi^{(k)}, L^{(k)})$ denote the outer-iteration value of the low-dimensional parameters.

1. **Initialize.** Choose starting values for $(\delta, \gamma_0, \alpha_0, \Pi, L)$, the number of importance-sampling draws R, and convergence tolerances for the Newton, inner L-BFGS, and outer L-BFGS steps. Draw

   $$z_r \sim \mathcal{N}(0, I_2), \qquad r = 1, \ldots, R,$$

   once and hold these base draws fixed throughout estimation. Using the same draws at every parameter value gives common random numbers and stabilizes the simulated objective.

2. **Compute Laplace proposals.** For each driver i, solve the posterior mode problem in (31), where the conditional log-likelihood $\ell_i(\nu)$ is defined in (30):

   $$\hat{\nu}_i^{(k)} \coloneqq \arg\max_{\nu} \left[\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2\right]$$

   by damped Newton iterations. These mode-finding problems are separable across drivers, so they can be solved in parallel over i. Starting from the previous outer iteration's mode, or from zero on the first iteration, iterate using the posterior gradient and negative Hessian in (32):

   $$\Delta\nu_i^{(k,n)} \coloneqq H_i(\nu_i^{(k,n)})^{-1} \nabla_\nu \left[\ell_i(\nu_i^{(k,n)}) - \tfrac{1}{2}\|\nu_i^{(k,n)}\|^2\right],$$

$$\nu_i^{(k,n+1)} := \nu_i^{(k,n)} + \lambda_i^{(k,n)} \Delta\nu_i^{(k,n)}, \qquad \lambda_i^{(k,n)} \in (0,1],$$

where $\lambda_i^{(k,n)}$ is a damping factor chosen by backtracking. At convergence, set

$$H_i^{(k)} := H_i(\hat{\nu}_i^{(k)}) = U_i^{(k)\prime} U_i^{(k)}.$$

3. **Generate proposal draws.** Using the common base draws and the Cholesky factors from step 2, form the driver-specific proposal draws

$$\nu_{ir}^{(k)} := \hat{\nu}_i^{(k)} + (U_i^{(k)})^{-1} z_r, \qquad r = 1, \ldots, R,$$

as in (33). These proposal draws are held fixed throughout outer iteration k.

4. **Compute importance weights and simulated likelihood.** For any candidate value of δ visited during the inner loop, evaluate the log importance weights, normalized importance weights, and simulated log-likelihood contributions using (34)—(36):

$$\log w_{ir}^{(k)}, \qquad a_{ir}^{(k)}, \qquad \log \hat{p}_i^{(k)}.$$

These quantities are then used to evaluate both the objective and the analytic gradients.

5. **Inner solve for δ.** Holding $(\gamma_0, \alpha_0^{(k)}, \Pi^{(k)}, L^{(k)})$ fixed, solve

$$\hat{\delta}^{(k)} \in \arg\min_{\delta} Q(\delta, \gamma_0, \alpha_0^{(k)}, \Pi^{(k)}, L^{(k)})$$

by L-BFGS using the objective in (27) and the gradient in (38). Each inner-loop objective and gradient evaluation calls step 4. This inner problem is convex in δ: for any fixed draw ν, each driver's conditional negative logit log-likelihood is a sum of log-sum-exp terms minus linear terms in δ, and the IV-GMM penalty is quadratic in δ. The exact marginal likelihood therefore remains log-concave in δ, and with the proposal draws fixed the simulated objective inherits the same convexity.

6. **Update γ_0.** Given the resulting $\hat{\delta}^{(k)}$, update γ_0 in closed form by minimizing the penalty implied by (28). Writing

$$P_Z := Z(Z'Z)^{-1}Z',$$

the update is

$$\hat{\gamma}_0^{(k)} = \arg\min_{\gamma} \tfrac{1}{2}(\hat{\delta}^{(k)} - \gamma\overline{D})' P_Z (\hat{\delta}^{(k)} - \gamma\overline{D}) = (\overline{D}' P_Z \overline{D})^{-1} \overline{D}' P_Z \hat{\delta}^{(k)}.$$

This is the IV coefficient from regressing the fitted mean utilities on market-level average duration using Z as instruments.

7. **Outer L-BFGS step.** Holding $(\hat{\delta}^{(k)}, \hat{\gamma}_0^{(k)})$ fixed, use the score formulas (39) and (40) to take

one outer L-BFGS step in (α_0, Π, L) and obtain

$$(\alpha_0^{(k+1)}, \Pi^{(k+1)}, L^{(k+1)}).$$

Because changing L changes the Laplace proposal, the next outer iteration recomputes the posterior modes, Hessians, proposal draws, and importance weights from scratch.

8. **Convergence test.** Stop when the outer L-BFGS iteration satisfies its convergence criterion; otherwise increment k and return to step 2.

At convergence, the reported estimator combines the profiled mean utilities $\hat{\delta}$, the IV update $\hat{\gamma}_0$, and the converged outer-loop parameters $(\hat{\alpha}_0, \hat{\Pi}, \hat{L})$.

C.3 Inference

We report robust standard errors for the low-dimensional parameter vector

$$\theta := (\gamma_0, \alpha_0, \mathrm{vec}(\Pi), \mathrm{vech}(L)),$$

treating the large vector of mean utilities δ as a nuisance parameter. Let H denote the full Hessian of the objective Q with respect to (δ, θ). The profiled Hessian for θ is the Schur complement

$$\tilde{H}_{\theta\theta} := H_{\theta\theta} - H_{\theta\delta} H_{\delta\delta}^{-1} H_{\delta\theta}.$$

The matrix $H_{\delta\delta}$ is large, but it has a simple structure. Stack the mean utilities as

$$\delta = (\delta_1', \ldots, \delta_{|\mathcal{M}|}')', \qquad \delta_m := (\delta_{mt})_{t \in \mathcal{T}_m}',$$

where $\mathcal{T}_m$ is the set of departure periods available in market m. Then we define the block diagonal matrix B as follows:

$$B = \mathrm{diag}(B_1, \ldots, B_{|\mathcal{M}|}).$$

For each driver i, draw r, and market m, define the market-m probability vector and observed-choice vector by

$$s_{i\ell,m}^{(r)} := \left(s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu_{ir})\right)_{t \in \mathcal{T}_m}, \qquad Y_{i\ell,m} := (Y_{i\ell t})_{t \in \mathcal{T}_m},$$

and the corresponding draw-specific score and Hessian contributions

$$G_{imr} := \sum_{\ell:\, m(i,\ell)=m} \left(Y_{i\ell,m} - s_{i\ell,m}^{(r)}\right), \qquad \Lambda_{imr} := \sum_{\ell:\, m(i,\ell)=m} \left[\mathrm{diag}(s_{i\ell,m}^{(r)}) - s_{i\ell,m}^{(r)} s_{i\ell,m}^{(r)\prime}\right].$$

Then $\partial \ell_i(\nu_{ir}) / \partial \delta_m = G_{imr}$ and

$$\frac{\partial^2 \ell_i(\nu_{ir})}{\partial \delta_m \partial \delta_m'} = -\Lambda_{imr}.$$

Differentiating (37) once more yields

$$\frac{\partial^2 \log \hat{p}_i}{\partial \delta_m \partial \delta_m'} = -\sum_{r=1}^{R} a_{ir} \Lambda_{imr} + \sum_{r=1}^{R} a_{ir} \left(G_{imr} - \bar{G}_{im}\right)\left(G_{imr} - \bar{G}_{im}\right)', \qquad \bar{G}_{im} \coloneqq \sum_{r=1}^{R} a_{ir} G_{imr}.$$

Accordingly, the market-m diagonal block is

$$B_m \coloneqq -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log \hat{p}_i}{\partial \delta_m \partial \delta_m'}.$$

In other words, each block B_m averages the market-m curvature of the simulated log-likelihood with respect to the vector of departure-time utilities within market m. With

$$C \coloneqq \frac{1}{N}(Z'Z)^{-1},$$

the full δ-δ Hessian is

$$H_{\delta\delta} = B + ZCZ',$$

where the penalty contributes the low-rank term $ZCZ' = \frac{1}{N} Z(Z'Z)^{-1}Z'$. We therefore evaluate $H_{\delta\delta}^{-1}$ using the Woodbury identity,

$$H_{\delta\delta}^{-1} = B^{-1} - B^{-1} Z \left(C^{-1} + Z' B^{-1} Z\right)^{-1} Z' B^{-1},$$

rather than forming a dense inverse. This reduces the computation to solving independent market-level systems involving B and one low-dimensional system involving the number of instruments (Grieco et al., 2025).

To construct the meat of the sandwich covariance, we profile both the likelihood scores and the penalty scores with respect to δ. For each driver i, let $(g_{i,\theta}, g_{i,\delta})$ denote the score contributions from the simulated log-likelihood. The corresponding profiled score is

$$g_{i,\delta} \coloneqq -\frac{1}{N} \nabla_\delta \log \hat{p}_i, \qquad g_{i,\theta} \coloneqq -\frac{1}{N} \begin{pmatrix} \partial \log \hat{p}_i / \partial \gamma_0 \\ \partial \log \hat{p}_i / \partial \alpha_0 \\ \partial \log \hat{p}_i / \partial \mathrm{vec}(\Pi) \\ \partial \log \hat{p}_i / \partial \mathrm{vech}(L) \end{pmatrix}.$$

Here $g_{i,\delta}$ is exactly the likelihood score appearing in (38) before adding the penalty term; the (α_0, Π) and L blocks are given by (39) and (40); and the γ_0 block follows from the same posterior-average-score identity (37):

$$\frac{\partial \log \hat{p}_i}{\partial \gamma_0} = \sum_{\ell \in \mathcal{L}_i} \sum_{t \in \mathcal{T}_i} \sum_{r=1}^{R} a_{ir} \left(Y_{i\ell t} - s_{i\ell t}(\delta, \gamma_0, \alpha_0, \Pi, L, \nu_{ir})\right) \left(\hat{D}_{i\ell t} - \overline{D}_{m(i,\ell)t}\right).$$

The corresponding profiled score is

$$\tilde{g}_{i,\theta} \coloneqq g_{i,\theta} - H_{\theta\delta} H_{\delta\delta}^{-1} g_{i,\delta}.$$

Likewise, let $c = 1, \ldots, C$ index the market-period observations entering the penalty, and let $\hat{V}_c$ and $\overline{D}_c$ denote the corresponding entries of $\hat{V}$ and $\overline{D}$. If e_c is the cth standard basis vector in the δ-space and e_γ is the selector for the γ_0 component of θ, define the penalty-score contributions by

$$g_{c,\delta}^{\text{pen}} \coloneqq \frac{1}{N} e_c (P_Z \hat{V})_c, \qquad g_{c,\theta}^{\text{pen}} \coloneqq -\frac{1}{N} \overline{D}_c (P_Z \hat{V})_c e_\gamma.$$

Summing these contributions over c recovers the full penalty gradients with respect to δ and γ_0; all other components of $g_{c,\theta}^{\text{pen}}$ are zero because the penalty depends only on (δ, γ_0). Their profiled versions are

$$\tilde{g}_{c,\theta}^{\text{pen}} \coloneqq g_{c,\theta}^{\text{pen}} - H_{\theta\delta} H_{\delta\delta}^{-1} g_{c,\delta}^{\text{pen}}.$$

The robust sandwich covariance estimator is then

$$\hat{V}_\theta = \tilde{H}_{\theta\theta}^{-1} \left(\sum_{i=1}^{N} \tilde{g}_{i,\theta} \tilde{g}_{i,\theta}' + \sum_{c=1}^{C} \tilde{g}_{c,\theta}^{\text{pen}} \left(\tilde{g}_{c,\theta}^{\text{pen}} \right)' \right) \tilde{H}_{\theta\theta}^{-1}. \tag{41}$$

This estimator allows for misspecification of the random-coefficient distribution and for the additional sampling variation induced by the IV-GMM penalty.

Finally, although we optimize over the Cholesky factor L, the covariance matrix $\Sigma = LL'$ is more interpretable. Accordingly, we use the delta method to obtain standard errors for

$$\vartheta \coloneqq (\gamma_0, \alpha_0, \text{vec}(\Pi), \text{vech}(\Sigma)).$$

If

$$J \coloneqq \frac{\partial \vartheta}{\partial \theta} = \begin{pmatrix} I & 0 \\ 0 & \partial \text{vech}(\Sigma) / \partial \text{vech}(L) \end{pmatrix},$$

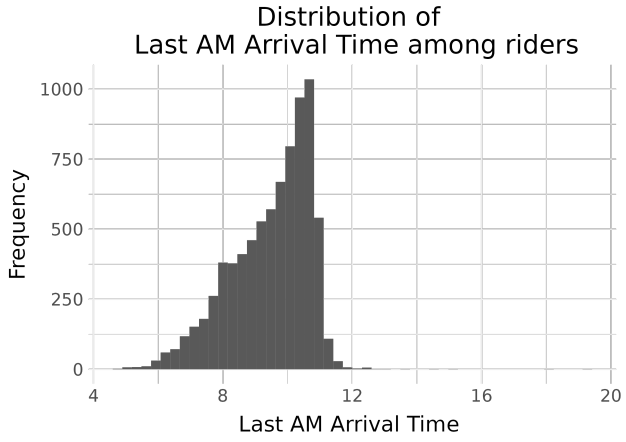
then the covariance matrix for the reported parameterization is

$$\hat{V}_\vartheta = J \hat{V}_\theta J'. \tag{42}$$

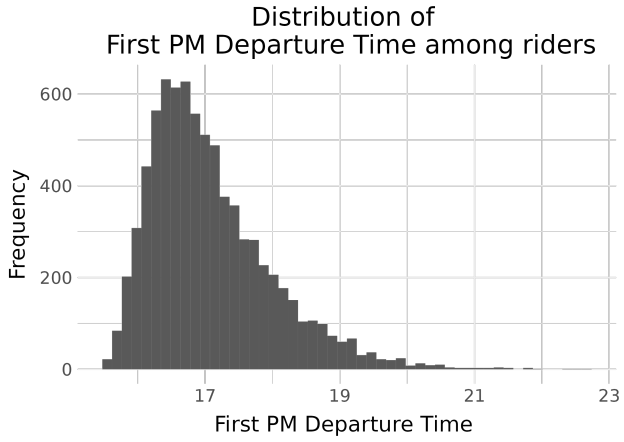
We use (41) and (42) to report the standard errors and confidence intervals in Table 3.

D Supplemental Tables and Figures

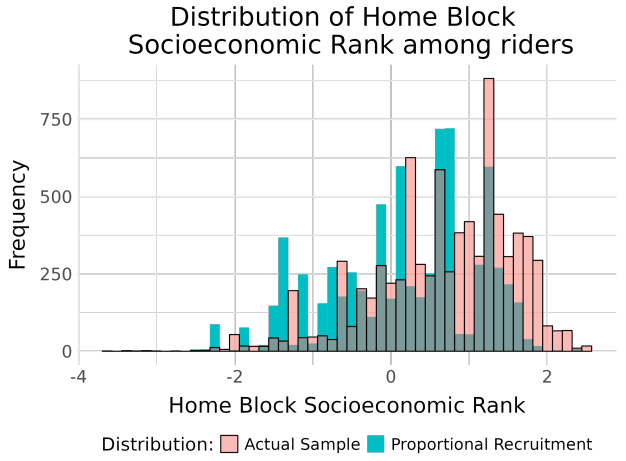
D.1 Distribution of Covariates



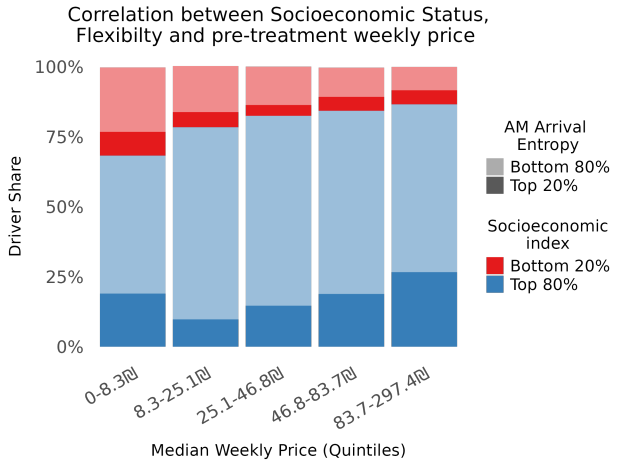
(a) Last AM Arrival Time Distribution



(b) First PM Departure Time Distribution



(c) Socioeconomic Index Distribution in Sample and in Population



(d) Relationship of socioeconomic status, flexibility, and exposure to pricing pre-treatment

Figure 16: Distribution of Covariates

D.2 Treatment Effect Heterogeneity

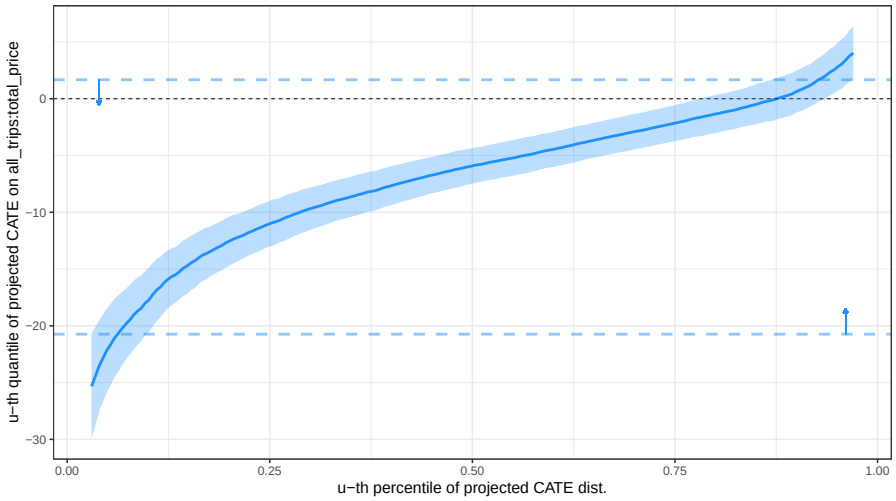


Figure 17: Estimated quantiles of the distribution of covariate-predicted driver-specific treatment effects, as well as a 90% uniform confidence band computed according to Chernozhukov et al. (2018). The dashed line with a downward-pointing arrow attached corresponds to the lower bound of the confidence interval for the largest quantile of the distribution of Δ_i, and the dashed line with an upward-pointing arrow attached corresponds to the upper bound of the confidence interval for the smallest quantile of the distribution of Δ_i; if the dashed line with the downward-pointing arrow attached is above the dashed line with the upward-pointing arrow attached, we can reject the null hypothesis of no heterogeneity in driver-specific treatment effects predicted by our covariates at the 10% level.

D.3 Demand Estimation Results

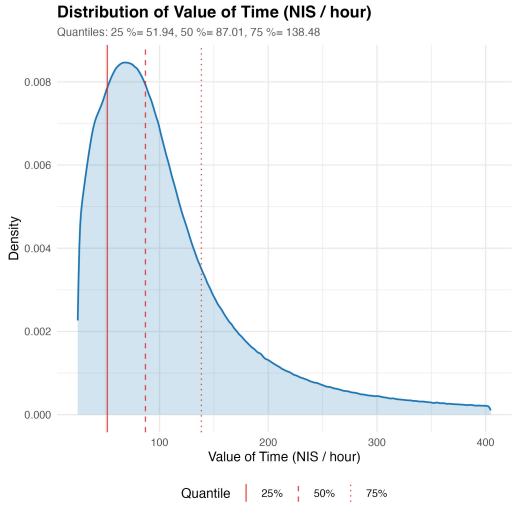


Figure 18: The estimated distribution of Value of Time γ_i/α_i.