Road Pricing under Heterogeneity*

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

August 10, 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: the extent of driver heterogeneity—both in preferences and in exposure to highly demanded roads and times of day—and the extent of revenue recycling. We quantify the impact of both forces using data from a large-scale experiment that tracked driving behavior with and without usage-based congestion pricing incentives for nearly 10,000 drivers over two years. Using a model of time-varying traffic equilibrium with heterogeneous drivers and a granular road network, we show that a congestion pricing policy using the experiment's price schedule would raise welfare by 93% of the value of adding a new lane to every highway and would raise enough revenue to fully compensate losers. Moreover, the extent of driver heterogeneity entails that optimal prices can improve net welfare with only a small fraction of toll revenue.

*Authors are listed alphabetically. This paper subsumes a previous manuscript titled, "Can Usage-Based Pricing Reduce Congestion?" by Ater, Ross, Shany, Turkel, and Vasserman. Eray Turkel contributed heavily to earlier stages of the project and we are very grateful for his many contributions. We thank Noa Bar-Nir, Fabian Barrett, Hadar Fuks, Shai Gilat, Mukun Liu, Andrew Meyers, Daniil Mikhailov, Yotam Nir, Freek van Sambeek, Guy Yakobovich, Doron Zamir, and Michelle Zheng for their exemplary research assistance. We are also grateful to our colleagues and numerous seminar and conference participants, and especially to our discussants Juan Camilo Castillo and Gabriel Kreindler for their helpful comments. We are grateful for financial support from the Israeli Science Foundation (541/22), The Israeli Smart Research Transportation Center, The Ministry of Science and Technology, and the Falk Institute for Economic Research. Computational Support was provided by the Data, Analytics, and Research Computing (DARC) group at the Stanford Graduate School of Business (RRID:SCR_022938).

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

‡The Wharton School, University of Pennsylvania. Email: bross1@wharton.upenn.edu

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

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

1 Introduction

Traffic congestion is a notorious source of inefficiency, frustration and environmental damage. Traffic delays are associated with billions of dollars of waste (Chandler-Wilde, 2023) and substantial greenhouse gas emissions (Barth and Boriboonsomsin, 2008), 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, Milan since 2008, and New York City since 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 recycled back to consumers, 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 cost of 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: 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 extent can the policy induce efficient sorting, and to what extent do drivers (and other individuals) internalize the benefits of the collected fees?

To quantify the importance of these factors in practice, we use original data from a usage-based congestion pricing experiment that tracked the driving behavior of nearly 10,000 drivers in Israel over the span of two years. Each driver in the experiment installed a monitoring device in their vehicle that recorded 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.

Our paper is composed of two main parts: an 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 individuals reduce their driving in congested times and places once they began to pay for their driving, and how? To answer these questions, we apply an event study approach, using Borusyak et al. (2024) to estimate dynamic average treatment effects across a number of driving behaviors in the four months after individuals begin to pay for their driving. Identification of treatment effects relies on the staggered recruitment of individuals to the experiment: an average of about 150 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 still being monitored without pricing who could serve as a control group to separate idiosyncrasies in driving behavior and time-varying aggregate driving conditions from the effects of the congestion pricing treatment.

Our treatment effect analysis suggests that drivers persistently decrease their driving in congested areas and times. Aggregating across post-treatment weeks, we find that drivers decrease a variety of measures of congested driving by an average of over 10% of a pre-treatment standard deviation, both on the extensive (e.g. number of trips) and the intensive (e.g. departure time) margins. We find substantial differences in the characteristics of drivers who responded most and least to the experimental pricing treatment on both margins. Using the Sorted Effects method (Chernozhukov et al., 2018), we find that the most responsive drivers in our sample exhibited more flexibility in their arrival and departure times and took more expensive trips during their "monitoring" period. Treatment effects also correlate with the availability of alternatives to driving: drivers with the biggest treatment effects on the number of driving trips they took per week tended to have better access to relevant public transit options.

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 on 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 segments at the same time, their expected trip durations increase according to the "congestion functions" of the road segments.

A key feature of our model is its focus on heterogeneity in demand for different parts of the road network. Drivers take trips defined by an origin, destination, and departure time window and are endowed with mixed logit preferences over transit convenience, travel 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 drivers' equilibrium choices. The congestion effects of overlaps depend on the physical characteristics of the roads (e.g., the number of lanes) 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 spread along the rest of 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 departure 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 trips taken by 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 estimate the 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. We then estimate the preference parameters that best rationalize the shares of departures in each 15 minute interval in 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, unobserved heterogeneity, 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 exposure to supply shocks: unusually heavy rainstorms across the country (The Times of Israel, 2020, 2021).

Our demand estimates show substantial heterogeneity in value of travel time (VOTT) across drivers: the median VOTT is close to the median hourly wage in Israel, but the 75th percentile is nearly 70% higher. This heterogeneity is correlated with demographics: lower income drivers have a higher disutility from price and a lower sensitivity to travel time. Preferences also vary with the availability of convenient public transportation: drivers with a shorter public transit alternative are less likely to drive—especially if they are lower income or live in denser areas. In order to accurately capture the interaction between preferences and the road network, we also estimate a model of supply of traffic flow. Using high frequency data from sensors along a major highway, we estimate the relationship between vehicle densities and traffic 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 cell phone location data.

Combining supply and demand, we solve for counterfactual traffic equilibria in Israel's largest commuting zone (the Central District around Tel Aviv) subject to different transportation network interventions. 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 about 15% (roughly 5 percentage points)—in the same range as the drops in traffic reported in the immediate aftermath of congestion pricing in London (Leape, 2006) and New York City (Badger et al., 2026). Diverting these drivers reduces travel times for trips that

are still taken, and with full revenue recycling the resulting gains would increase average welfare.

To put the magnitude of these welfare gains in perspective, we benchmark congestion pricing against two of the most commonly proposed alternatives for relieving congestion: physically expanding road capacity and improving public transit. We find that congestion pricing using the experiment's price schedule would deliver a net welfare gain equal to 93% of the gain from adding a lane to every highway in the Central District, before accounting for construction costs. Expanding transit based on the planned Tel Aviv metro system, which is scheduled to open in 2037, would deliver substantially less on net: only 8% of trips would have a faster transit alternative and equilibrium driving probabilities would decrease by less than 1%. This exercise is conservative: it doesn't account for general equilibrium effects like changes in residential and workplace choices that might result from a new subway network. Still, the difference in magnitudes is informative about the relative gains from congestion pricing, which require substantially smaller investments in infrastructure.

To assess the full scope for welfare gains, we compute the welfare-maximizing prices under the experiment's price schedule. We find that the optimal usage based price schedule would charge substantially more for travel in the central metropolis (metro) zone during off-peak hours and in the periphery during peak hours than the experiment's price schedule. These prices would increase welfare by 2 NIS per trip-day—a 17% improvement over adding a lane to every highway. A welfare-maximizing cordon-based price schedule would deliver a substantially smaller gain: only 0.48 NIS per trip-day. This reflects the lower ability of cordon pricing to efficiently target heavy users of congested roads without harming drivers who just pass through briefly.

Our counterfactuals make clear that both forces highlighted by our framework—heterogeneity and revenue recycling—are quantitatively important. Across every pricing scheme, the average driver's raw consumer surplus (before any rebating) falls, and the gains accrue disproportionately to drivers with the highest values of time. The revenues collected are instrumental to generating large welfare gains, and, if distributed appropriately, can generate a Pareto improvement. However, revenues are not strictly necessary: even with no revenue recycling at all, an optimally targeted usage-based pricing schedule can modestly improve net welfare. As our theory suggests, this last result depends entirely on heterogeneity. When we shut down heterogeneity in preferences or in exposure to the road network, positive prices can no longer improve welfare on their own—at least 44% of revenues would need to be recycled under homogeneous preferences, and at least 21% under a homogeneous road network.

Together, our results suggest that the levels of congestion, values of time, and heterogeneity across drivers in Israel are so high that even simple congestion prices can generate substantial welfare gains. To get the most out of pricing, the price schedule should be designed not only to price the most congested roads and times, but also to target drivers whose trips contribute most to congestion on net and who have the most flexibility to respond. Our framework can evaluate any such pricing scheme or other network intervention, and can be applied to any setting where comparable data is available.

1.1 Related Work

Our paper builds on foundational theoretical work on the use of pricing policies for relieving congestion externalities (Pigou, 1920; Knight, 1924; Walters, 1961; Vickrey, 1969; Henderson, 1974; Small, 1982;

Arnott et al., 1993, 1994; Fosgerau and de Palma, 2013; van den Berg and Verhoef, 2011; Hall, 2018; Dipasquale et al., 2026) as well as a rich empirical literature quantifying the economic costs of traffic congestion (Small et al., 2005; Akbar and Duranton, 2017; Couture et al., 2018; Yang et al., 2020; Hall, 2021; Akbar et al., 2023; Tarduno, 2024; Cook and Li, 2025; Small et al., 2024).

We contribute experimental evidence on the effectiveness of congestion pricing policies to induce behavioral change and improve overall welfare. Previous work has shown mixed results: experiments in Bangalore (Kreindler, 2024) and Melbourne (Martin and Thornton, 2017) find that congestion prices on driving induce modest behavioral responses unlikely to substantially reduce congestion. Hintermann et al. (2026) find that prices can induce drivers with access to high quality public transit in Switzerland to reduce the social cost of their travel by 4.6% in partial equilibrium, although they do not consider induced demand effects. Our study considers the largest field experiment to date with nearly 10,000 participants and a congestion pricing treatment that lasted over a year. As in many congested cities in the United States, the vast majority of participants in our experiment did not have rail-based public transit alternatives for their driving trips. Furthermore, while previous experiments concluded prior to the COVID-19 pandemic, our experiment largely took place during and after the COVID-19 recovery and accounts for persistent changes in workplace flexibility due to improvements in work from home technology during the pandemic. Perhaps in part due to these differences, we find a larger effect on trip demand from congestion pricing.

Our paper relates most closely to a recent literature applying techniques from empirical industrial organization to assess the viability of congestion pricing policies in large metropolitan areas. Using data from his field experiment in Bangalore, Kreindler (2024) simulates a model of city-wide congestion in which drivers with trips of different lengths choose their departure times and proceed at slower speeds if they overlap with larger masses of other drivers at the same time. Durrmeyer and Martinez (2022) and Almagro et al. (2024) use cross-sectional mode share data to simulate models of mode choice in which drivers commuting between neighborhoods in Paris or Chicago choose between driving, public transit or other alternatives. Durrmeyer and Martinez (2022) estimate that drivers in Paris are more elastic in their mode choice than in their choice to travel during peak or off-peak hours and show that per-kilometer usage-based prices are nearly as effective as trip-based personalized prices. Almagro et al. (2024) show that the ways in which toll revenues are used can be critical to the value proposition of congestion pricing: revenue rebates alone would be not be enough to improve consumer surplus through per-mile prices in Chicago, but optimal use of revenues to improve public transit would be.

Our paper uses panel variation to estimate the elasticity of time-shifting in 15 minute increments and mode-shifting for the same trip taken by the same driver. Although we do not observe mode choice directly and bundle all non-driving choices into a single category, we find that drivers with more attractive public transit alternatives for the same trip are more likely to substitute away from driving upon a price increase. All three related papers find that the welfare benefits from congestion reduction alone cannot justify a road pricing policy and that revenues must be recycled for such a policy to be net beneficial. Our theoretical framework suggests that this may be a result of the coarseness of driver and network heterogeneity in these studies: in order for a road pricing policy to increase average welfare before revenue recycling, it must be able to not only reduce the load on congested roads but

also to efficiently sort drivers with sufficiently different values for time and money. Empirically, we find that, as in these studies, the congestion prices that optimize total welfare including revenues generate a decrease in raw consumer surplus that is compensated by a larger increase in revenues that can be redistributed. However, we also find that it is possible to choose prices that improve raw consumer surplus with no revenue recycling at all.

Our model holds residential and workplace locations fixed and focuses on the shorter-run equilibrium of trip participation, departure times, and travel times, matching the horizon of our experiment. Other papers instead study how congestion policies affect agglomeration, residential sorting, and the locations of households and firms (Brinkman, 2016; Allen and Arkolakis, 2022; Herzog, 2024; Barwick et al., 2024; Hierons, 2025; Bordeu, 2025).

Finally, our comparison of road pricing with road and transit infrastructure investments relates to a broader literature on evaluating the impacts of transportation infrastructure improvements (Fogel, 1964; Duranton and Turner, 2011; Billings, 2011; Anderson, 2014; Donaldson, 2018) and quantitative spatial models (Fajgelbaum and Schaal, 2020; Allen and Arkolakis, 2022; Tsivanidis, 2026; Severen, 2023; Bordeu, 2025); see Donaldson (2025); Allen et al. (2025); Severen (2026) for recent overviews. While we cannot speak to the general equilibrium implications of infrastructure improvements, we model the transportation network at the level of individual road segments and are thus able to evaluate specific infrastructure improvements like adding lanes to highways or building specific subway lines, in the spirit of papers like Kreindler et al. (2023) and Brancaccio et al. (2026).

2 Motivating Example

In this section, we present a minimal illustrative example to demonstrate the roles of driver heterogeneity and revenue recycling 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:

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

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 $\Lambda = 0$, the toll revenue is kept entirely by the planner and each driver bears the full cost of the toll with no rebate. When $\Lambda = 1$, the planner returns all of the toll revenue back to drivers (e.g., in the form of a lump sum rebate). Intermediate values of Λ between 0 and 1 correspond to partial revenue recycling—for instance, if a portion of the toll revenue is used to fund the operation and enforcement of the toll. It is also possible for Λ to be greater than 1 if the planner is able to use the toll revenue to fund public goods with economies of scale, such as public transportation.

Efficiency Gap with Homogeneous Types When all drivers have the same preferences (up to their idiosyncratic taste shocks) and tolls are fully recycled ($\Lambda = 1$), 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 Toll Revenues The familiar result above relies crucially on the assumption that the toll revenue is fully returned 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 constitute 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.

The logic for this argument follows a similar intuition to that in Vickrey (1969). In equilibrium, the marginal driver must always be indifferent between driving and not driving. A toll therefore switches the marginal driver from paying in "congestion" to paying a commensurate amount in money. If drivers are homogeneous, then inframarginal drivers are differentiated only by their taste shock and value the toll cost and benefit in the same way as marginal drivers. To reduce the number of drivers on the road, the toll must therefore reduce the overall value of driving for all drivers, making drivers who continue to drive strictly worse off.

Notably, this argument does not depend on the convexity of the duration function $D(\cdot)$ or the distribution of the taste shocks $F(\cdot)$. For any toll $p > 0$, we can subtract the equilibrium indifference conditions in Equation (1) to parametrize the congestion relief that can be sustained in equilibrium as:

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

Since $F^{-1}(q^*(p)) < F^{-1}(q^*(0))$, the change in the threshold for driving over the outside option 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 difference in the marginal taste shock for driving. Meanwhile, marginal drivers who exit are 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 differ through their willingness to pay for

travel time, represented by 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 $q^*(0) = F\big(V - \gamma D(q^*(0))\big)$. But once a toll is imposed, types separate. A driver with high α_i (high price sensitivity and therefore low willingness to pay) reduces their driving more than a driver with low α_i (low price sensitivity, high willingness to pay). The toll thus reallocates trips away from those who value them least in money terms and toward those who value them most.

More generally, for any Λ, there is a welfare improving price $p > 0$ whenever the derivative of $W_\Lambda(p)$ with respect to p is positive at $p = 0$:[1]

$$\left.\frac{\mathrm{d}W_\Lambda}{\mathrm{d}p}\right|_{p=0} = \underbrace{-\dot{D}(0)q^*(0)\bar{\phi}}_{\text{average congestion relief}} + \underbrace{-\dot{D}(0)\,\mathrm{Cov}(q_i^*(0),\,\phi_i)}_{\text{sorting gain}} - \underbrace{(1-\Lambda)q^*(0)}_{\text{revenue leakage}}.$$

Here, $\phi_i = \frac{\gamma_i}{\alpha_i}$ is the value of time for driver i with population average $\bar{\phi} \equiv \sum_i \omega_i(\gamma_i/\alpha_i)$, and $\dot{D}(0)$ is the total derivative of duration with respect to a marginal toll at $p = 0$. Since $\dot{D}(0) < 0$, the first term is the overall average value of time saved due to a marginal toll at $p = 0$. The second term is the sorting gain: the extent to which drivers who take the trip (and so benefit from the faster road) have higher values of time than those who do not. The third term is the revenue leakage: the total toll amount that is extracted from drivers and not recycled. Dividing by $q^*(0)$, the condition becomes:

$$\left|\dot{D}(0)\right|\left[\bar{\phi} + \mathrm{Cov}\left(\frac{q_i^*(0)}{q^*(0)},\,\phi_i\right)\right] > 1 - \Lambda. \tag{2}$$

That is, a marginal toll is welfare-improving only when the value of time saved, weighted by the relative probability of driving, is large enough to offset each dollar of unrecycled toll payments.

Heterogeneity in Network Exposure The sorting mechanism above relies on drivers differing in preferences, summarized by ϕ_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, as illustrated in Figure 1. 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.

Because all drivers must pass through segment 2 to reach their destinations, a toll p_2 on segment 2 can only be welfare-improving if drivers are heterogeneous or if some of the toll revenue is recycled, as in the one segment case. However, a toll p_1 on segment 1 provides an additional sorting mechanism: it screens out some of the m_{12} through-drivers off the road, which reduces congestion on segment 2 for the m_2 drivers without charging them. If the duration function $D_2(\cdot)$ on segment 2 is sufficiently convex and the common value of time $\phi = \gamma/\alpha$ is sufficiently high, then the travel-time savings enjoyed by the m_2 segment-2 entrants may exceed the toll burden borne by the m_{12} segment-1 entrants, generating a welfare improvement. More generally, the potential for welfare-improving tolls grows as the network

[1] See Appendix Section B for a derivation.

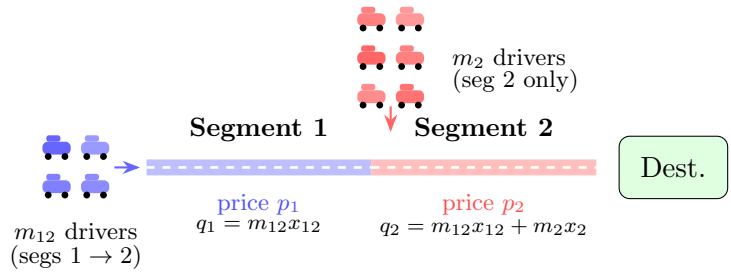


Figure 1: Heterogeneity in network exposure with two sequential segments. Through-drivers (mass m_{12}) traverse both segments to reach their destination and are exposed to both tolls p_1 and p_2, while segment-2-only drivers (mass m_2) enter directly at segment 2 and bear only p_2. Even when all drivers are identical in (α, γ, V), a toll p_1 on segment 1 screens some through-drivers off the road, relieving congestion on segment 2 for the m_2 drivers who pay none of the p_1 burden.

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. In Section 4.1, we show that drivers exhibit substantial heterogeneity in price sensitivity and time preferences. In Section 5, we expand our motivating example into a full model of network equilibrium with heterogeneous drivers. In Section 6, we combine these two elements to estimate a model of demand for travel and the supply of road capacity across the largest commuting zone in Israel. In Section 7, we combine supply and demand to estimate the equilibrium welfare effects of congestion pricing under different pricing schemes and assess the level of revenue recycling 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.[2] Since late 2019, we have served as an independent research team focused on the economic evaluation of the experiment.

Approximately 10,000 drivers were recruited to participate in the experiment between January 2020 and October 2020. Recruitment was stratified by geographic region to capture the relative contribution to congestion from the driving behavior of each area, and staggered over time with a mean (median) of 648 (555) drivers each month. The majority of participants resided in the broader Tel Aviv commuting zone, which is home to nearly half of the Israeli population. Nearly all participants continued with the experiment until its conclusion in November 2021.

[2] Ayalon Highways Co. is a government entitiy similar to a national highway administration that is responsible for technology-driven traffic management, including smart traffic lights, managed lanes, and vehicle-related pilot programs.

The operation of the experiment was managed by two companies: Pango, Israel's leading parking payment company, and Pointer, an Auto-Tech company that specializes in the protection and location of stolen vehicles. Recruitment combined mass-media public outreach with targeted marketing. Backed by the Israeli Ministries of Finance and Transportation and Ayalon Highways Co., a broad national campaign using billboards, digital advertising, and mass media invited the general public to volunteer via self-registration. Simultaneously, the two operating companies—Pango and Pointer—conducted targeted marketing campaigns directed at their existing customer bases. Each company then oversaw the full scope of experiment participation for its recruits. Upon recruitment, drivers underwent a brief telephone interview, during which they were asked to provide administrative information (e.g., documentation for their bank account, driver's license, and vehicle ownership) and were given a brief explanation of the experiment's rules. After signing a participation agreement, drivers were instructed to schedule the installation of a monitoring instrument: an inconspicuous, difficult-to-remove device that tracks whether a vehicle's engine is running or idle and transmits its GPS location. Most drivers already had a qualifying GPS tracking device installed in their vehicle as part of insurance requirements. In such cases, there was no need to install a second one.

During the first six months after a driver joined the experiment, their driving behavior was monitored and recorded without communication or 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.[3]

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." Only high-traffic driving was priced: during peak hours, kilometers driven in the metro were charged 1.5 NIS per kilometer and kilometers driven in the metro fringe were charged 0.3 NIS per kilometer. During off-peak hours, driving in the metro was charged 0.1 NIS per kilometer, while all other driving was free.

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 week of recruitment, which started on 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

[3]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.

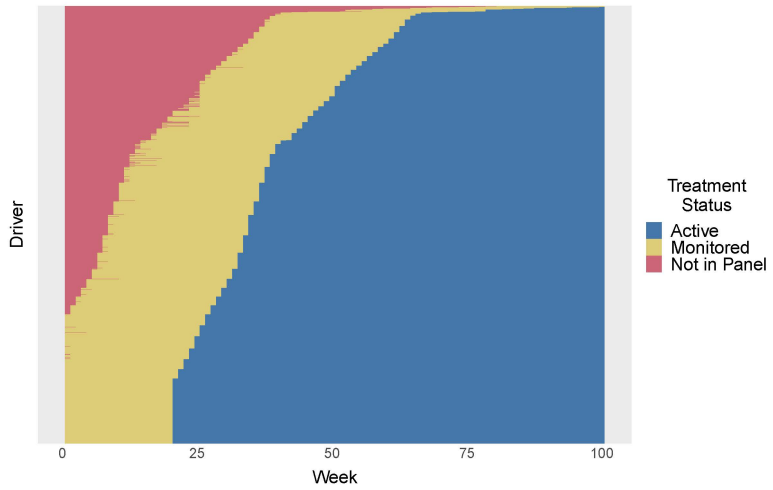


Figure 2: Participant status by week, starting on the week of January 5th, 2020. The y axis indexes drivers eventually recruited for our study, and the x axis indexes weeks relative to the first week of recruitment, which was January 5th, 2020. The coordinate corresponding to a given driver and week 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.

of a Traffic Analysis Zone (TAZ), the distance traveled, the trip duration, and the price that the driver paid (if they were in active status) or would have paid (if they were in monitoring status), based on the per-kilometer fee structure. Similar to US Census blocks, TAZs are non-overlapping geographic polygons with similar numbers of inhabitants that are designed to facilitate 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. In addition, while we cannot see the exact route that drivers take, we do see a breakdown of each 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.

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 their 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;[4] and the pollution level of the car and other car characteristics based on the vehicle's registration information.

3.2 Auxiliary Data

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

Highway Sensor Data We use data on average traffic speeds and volumes collected at 5-minute intervals along the Ayalon Highway to estimate the relationship between traffic densities and flows

[4] A dedicated mobile app was designed for the experiment to provide drivers with information on their historical spending, remaining budget, and estimated times and prices for hypothetical trips. Drivers were invited to download the app once they switched to the "active" status.

(e.g. congestion functions) on different roads in Section 6.2. The Ayalon Highway is the main highway that traverses the Tel Aviv Metropolitan Area and is operated by Netivey Ayalon Company. Our data was collected using Inductive Loop Detectors (ILD) sensors at different locations along the Ayalon Highway throughout 2020 and 2021. See Appendix Figure F1 for sensor locations.

Open Street Map 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 reconstruct the route that each trip takes through the road network at the level of individual road segments, following the sequence of TAZs that the trip traverses. 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 hold each trip's route fixed when considering how changes in road prices and speeds affect driving decisions. Appendix Section D describes how we reconstruct these routes on the OSM network and impute their durations and prices.

GTFS Bus GPS Pings We use high-frequency GPS pings from the Israeli public bus network to infer real-time speeds on every road segment at every date and time in our panel. The raw data consist of 1.4 billion GPS pings from 925 bus lines, each recording a bus's location, timestamp, and trip identifier. We match official transit-route shape files from the Israeli National Authority for Public Transport (NAPT) to the OSM road network and use consecutive pings to estimate bus speeds on the traversed segments. Because direct estimates cover only a subset of segment–time cells, we impute the remaining speeds using a sequence of fixed-effects regressions. This procedure yields a panel of bus-speed estimates for every OSM road segment, 15-minute time bin, and day in the sample, which we use to impute counterfactual trip durations at different departure times. See Appendix Section D.2 for a full description of this procedure.

Public Transit Data We use several sources of data on public transit availability and usage. First, we use information on historical public transit schedules and bus stop locations to calculate possible public transit alternatives for the observed trips conducted by the drivers, as described in more detail in Section 3.4. This data is published and made publicly available by the Israeli Ministry of Transportation on data.gov.il in the General Transit Feed Specification (GTFS) format. We construct each trip's public-transit alternative—its feasible itinerary, travel time, modal composition, and fare—by routing the observed origin, destination, and departure time through the OpenTripPlanner engine, using the Israeli GTFS schedule together with the OpenStreetMap road and pedestrian network. See Appendix Section D.5 for a full description of this procedure.

Cell Phone Trip Flows To reweight trips in our data to be representative of population-level origin-destination flows (see Appendix Section F.3), we use hourly origin-destination flows built from cell phone location data covering all of Israel in 2018 and 2019, collected by Israel's Ministry of Transportation and Road Safety as part of an update to their national travel demand model (Cohen et al., 2021). The data draw on antenna-triangulated location records from an average of 3.67 million monthly users of Israel's two largest cellular carriers—roughly half of Israel's population aged 8 and

above. These location records are aggregated into average counts across 2018 and 2019 weekdays of trips between every pair of 250 coarser groups of TAZs in each hour of the day.

Demographic Information We use demographic information obtained from the Israeli Central Bureau of Statistics (CBS) for each TAZ. This includes information on population density and growth and an index of socioeconomic status, both of which are used throughout our analysis.[5] We do not observe administrative data on income or employment for the drivers in the experiment. As such, we impute demographic information on the basis of each driver's inferred home TAZ.[6]

3.3 Habitual Trips

Much of our analysis focuses on *habitual trips*: recurring origin–destination routes that a driver takes regularly enough to reflect a stable underlying travel need, such as a commute. We define a habitual trip as a route between the same origin and destination that a given driver takes within a two-hour window on at least three occasions over a rolling eight-week period. Applying this procedure to our 23.4 million raw driving records yields 134,672 habitual trips (corresponding to 1,191,202 realized trip instances), with at least one habitual trip identified for 10,073 of the 10,441 drivers in our sample. The panel of habitual trips allows us to differentiate between intensive (e.g., time-shifting) and extensive (e.g., driving less) responses to the experimental treatment and forms the basis of the demand model that we estimate in Section 6. Appendix Section D details the full construction of the habitual trip panel, including trip-specific imputations for routes, durations, prices, and public-transit alternatives.

3.4 Descriptive Statistics

Baseline Driving The average driver in our experimental panel took 14 trips per work-week (Sunday through Thursday in Israel, excluding holidays) during their monitoring period, covering 188 kilometers over the course of 6.2 hours behind the wheel. The majority of the kilometers driven and time spent driving was on cross-city trips—the most common types of trips for commuting. Cross-city trips also constitute the majority of driving in congested zones, as designated by the experiment's per-Km fee schedule: if they had been paying for these trips, the average driver would have incurred 46 NIS in total congestion fees per week, two-thirds of which derived from cross-city driving.

Appendix Table A1 presents a detailed breakdown of trip counts, durations, speeds and prices across the distribution of drivers. About half of within-city trips have distinct origins and destinations ("non-cycle"). The other half, "cycle" trips, have the same origin and destination TAZ. 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. While the average "cycle" trip has a similar distance, duration

[5] The socioeconomic index is an official statistic published by the CBS to reflect differences in economic circumstances across different regions of Israel. The index is calculated on the basis of several variables, including the rate of public funding dependency, average years of schooling in the adult population, percentage of academic degree holders, levels of employment and household income. Lower values correspond to lower socioeconomic status.

[6] 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 inferred to be the driver's home TAZ.

and price to the average "non-cycle" trip, the latter contributes substantially more to total driving and congestion exposure at higher quantiles.

Because drivers in our experiment were recruited on the basis of their contribution to congestion (and were required to have a personal vehicle available for regular use), the average driver in our sample is not representative of the general population of Israel. Compared to the general population, the distribution of home TAZs in our sample under-represents the bottom quartile of socioeconomic status and over-represents the top quartile.[7] Compared to average driving statistics across the country, drivers in our sample take more trips per week, and drive fewer but slower kilometers than the average Israeli driver.[8] However driving patterns in our sample are similar to those in major metropolitan areas of the United States. US drivers took about 11.4 trips per work week on average in 2022 (down from 16.85 in 2017) (U.S. Department of Transportation, Federal Highway Administration, 2017). Among these, the average commute time was 25.7 minutes nationally, but over 30 minutes in major metropolitan areas like New York City and Washington DC (U.S. Census Bureau, 2022).

4 Causal Effects of Pricing

The staggered recruitment of drivers into the experiment facilitates an event study design to evaluate the average impact of pricing on the driving behavior of participants. We apply the difference-in-differences framework of Borusyak et al. (2024) to estimate cohort average treatment effects on the treated (CATTs) τ_{ct} for each cohort of drivers c, defined by the week in which they switched from "monitoring" to "active" status, and each calendar week t after their activation date. For each week t, we compute potential outcomes on the basis of the drivers who are still in the "monitoring" phase, subject to a two-way fixed effects parametrization. We then generate CATT estimates $\hat{\tau}_{ct}$ and aggregates by relative time to treatment $e = t - c$, by averaging the difference between observed outcomes and the potential outcomes predicted by the two-way fixed effects model. See Appendix Section C.1 for details on identification and estimation.

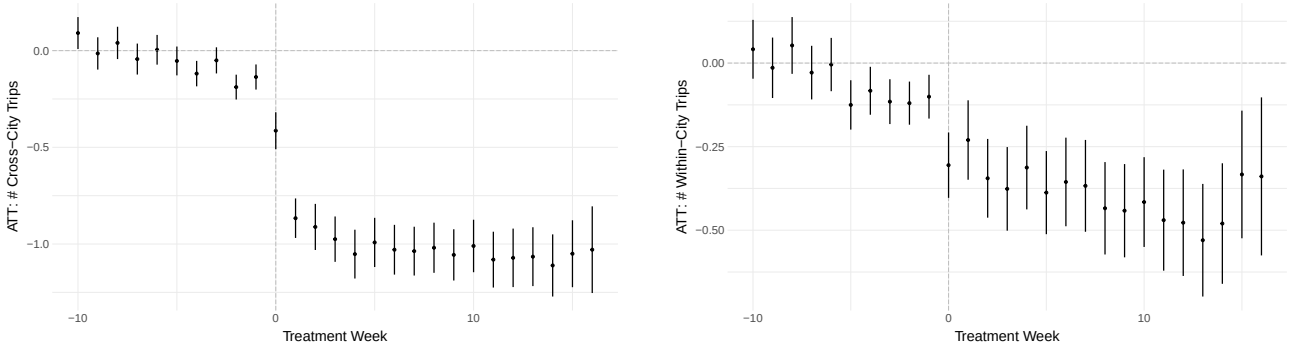
4.1 Average Treatment Effects

Figure 3 presents average treatment effect estimates for the number of trips taken over the 10 weeks before and 16 weeks after each driver's treatment week, along with associated 95% simultaneous confidence bands. Because different types of trips had different levels of exposure to pricing, we present separate event studies for cross-city trips and within-city non-loop (e.g., with different origin and destination points) trips—about 20% of the pre-treatment mean. As Figure 3a 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.[9]

[7] See Appendix Figure A1 for a comparison of the distribution of socioeconomic status in our sample to the population.

[8] Statistics on the general population are based on the most recent available cellular survey data. The average Israeli driver takes 1.9 trips per weekday and covers 21 Km per trip, spending less than one hour a day on the road.

[9] 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



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

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

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

Appendix Section A.2 presents corresponding event studies for the total weekly price, price per trip, and departure times for cross-city and within-city non-loop trips. The total weekly price of cross-city trips decreased by about 7.5 NIS per week on average, while the price per cross-city trip taken decreased by about 0.5 NIS per trip on average. This suggests that the nature of the trips that drivers took changed in response to the pricing treatment: they reduced their driving on the most expensive trips, changed the timing or routing of their trips to avoid peak pricing, or a combination of both.[10]

As Figure 3b shows, the treatment response for within-city non-loop trips was more gradual, with an average decrease of 0.25-0.5 trips per week. However the price dynamics were different: the average price per trip taken increased slightly over the course of the 16 weeks after treatment, while the total weekly price increased initially and then stabilized at zero. Appendix Section A.2 also presents event studies for the number and total price of within-city loop trips. Neither the number of trips nor the total weekly price of within-city loop trips changed significantly after treatment.

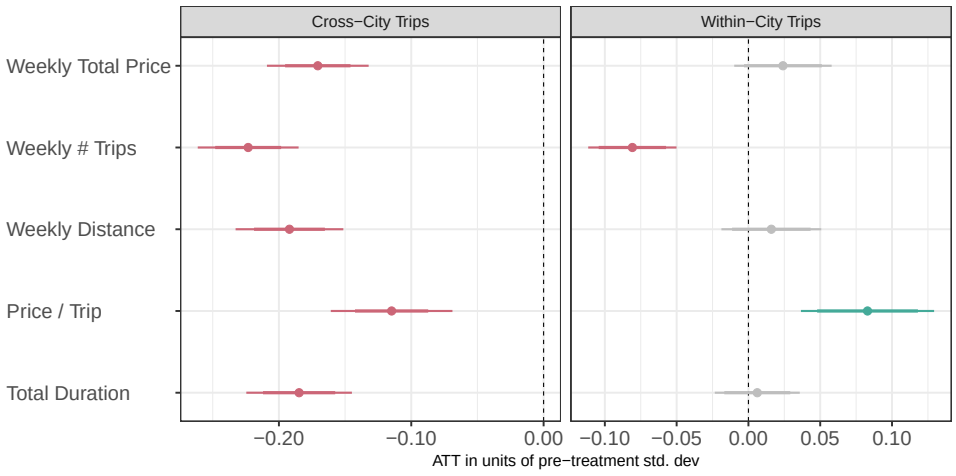


Figure 4: Average Treatment Effects Across Congestion-Related Outcomes

treatment week is set as two weeks prior to their measured treatment week (Callaway and Sant'Anna, 2021).

[10] Appendix Figure A5a shows that the average first departure time for evening cross-city trips (for which peak hours spanned 3:30 to 6:30PM) became later over the course of the 16 weeks after treatment.

Figure 4 summarizes the treatment effects across different measures of congestion-related driving for cross-city and within-city trips. The average treatment effect on the treated (ATT) for 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 3, 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. We plot the average treatment effect estimates for the other measures in the figure in the same way, normalizing the ATT estimate for each measure to the units of its pre-treatment standard deviation.[11]

As the left panel of Figure 4 shows, the treatment response for all of the measures of congestion-related driving on cross-city trips is negative in similar relative magnitudes—between 0.1 and 0.2 pre-treatment standard deviations. By contrast, there was no substantial impact on aggregate congestion-related driving 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 weekly price, total weekly distance or total weekly duration 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: timing and routing. Zooming in on Cross-City trips, Figure 5 plots the average treatment effects on measures related to the time of departure for the first trip of the 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.

[11] Because the measures 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 measures 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%. 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). 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.

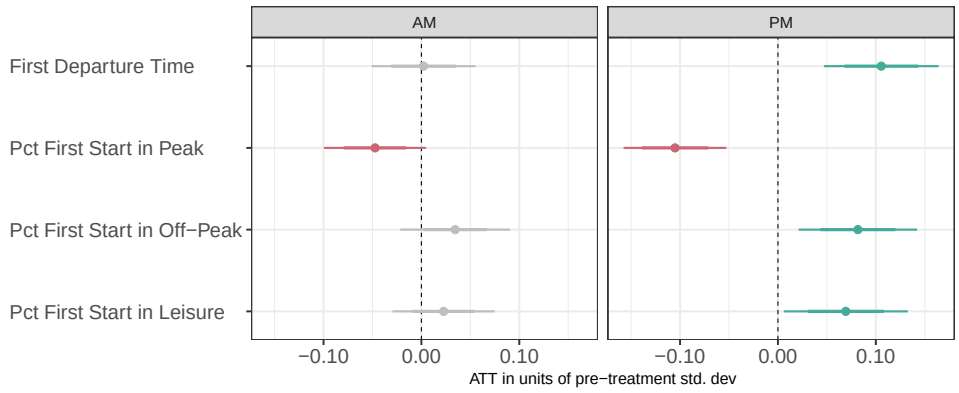


Figure 5: Average Treatment Effects on Departure Times for Cross-City Trips

Figure 6 plots the average treatment effects on the share of each trip that is spent in each of our nine pricing buckets, among PM cross-city trips. Although only one 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 off-peak and leisure times. At the same time, the share of trip distances covered in fringe and periphery areas during peak times did not change significantly (or even shifted down a bit in the case of fringe areas). 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—not where they drove. This is consistent with the results in Figure 5, which show that drivers shifted their start time for driving away from peak hours. It also suggests that drivers did not respond by changing their routes in a substantive way.

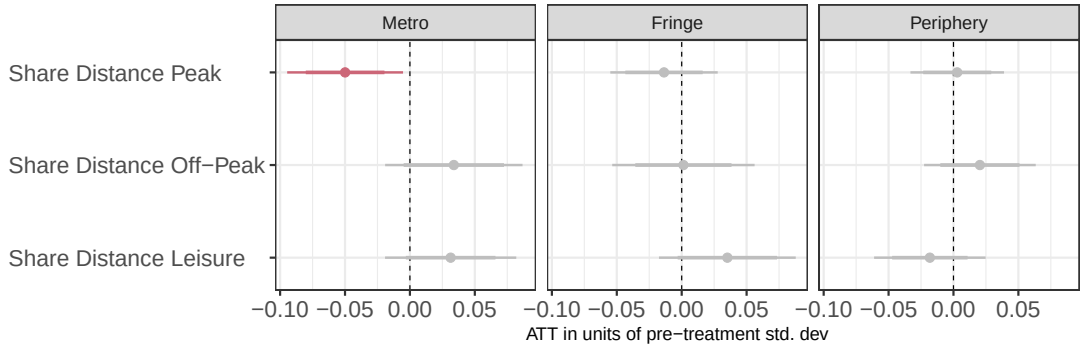


Figure 6: Average Treatment Effects on Trip Shares by Price Bucket for PM Cross-City Trips

4.2 Heterogeneity in Treatment Effects

The average treatment effects in Section 4.1 aggregate over substantial heterogeneity across drivers with different baseline driving needs, flexibility, and socioeconomic status. To characterize the nature and degree of heterogeneity across different dimensions of driving behavior, we apply the Sorted Effects method developed in Chernozhukov et al. (2018). For each outcome of interest, we compare the

average characteristics of 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. See Appendix Section C.2 for details.[12]

Figure 7 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 the baseline.

In some ways, it is intuitive that the 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 7 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 discernible difference in socioeconomic status between the most and least responsive drivers.

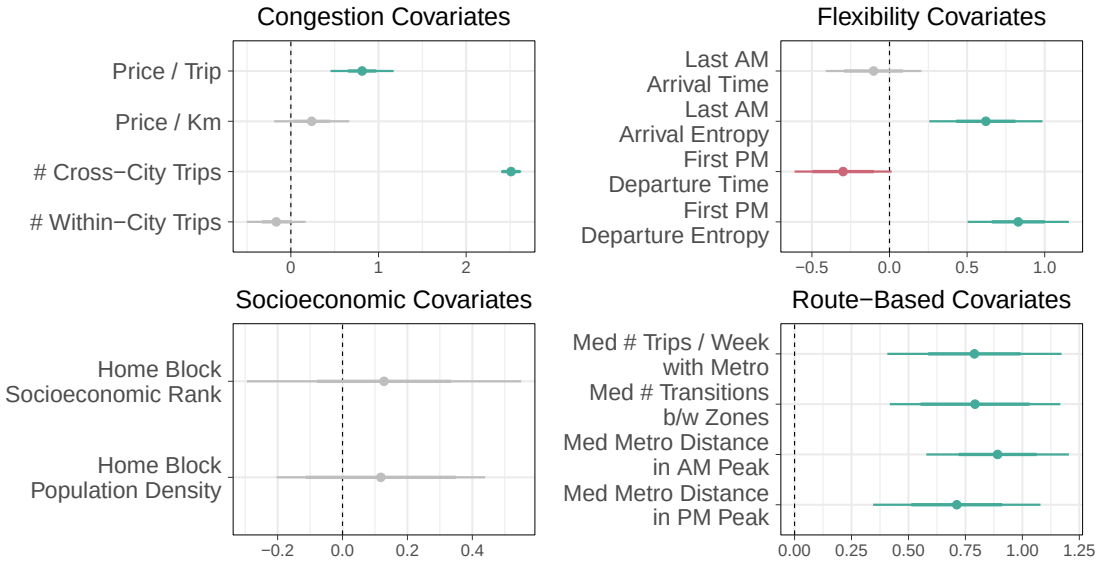


Figure 7: Sorted effects results for treatment effects on the number of cross-city trips taken per week. The x-axis shows the average difference in covariates, measured in standard-deviation units, between riders with the most positive CATEs (top 20%) and those with the most negative CATEs (bottom 20%) for the outcome.

[12] See Appendix Table A2 for a summary of the covariates used to measure heterogeneity in this exercise.

Figure 8 plots the sorted effects results for treatment responses in the first evening departure time for cross-city trips. As we discussed in Section 4.1, a stronger treatment response in this context corresponds to a bigger shift to a later first departure time. As with the total number of trips, 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 over-all 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 either the rush-hour trip was delayed (and any later trips were delayed as well) or that the rush-hour trip was replaced all together.

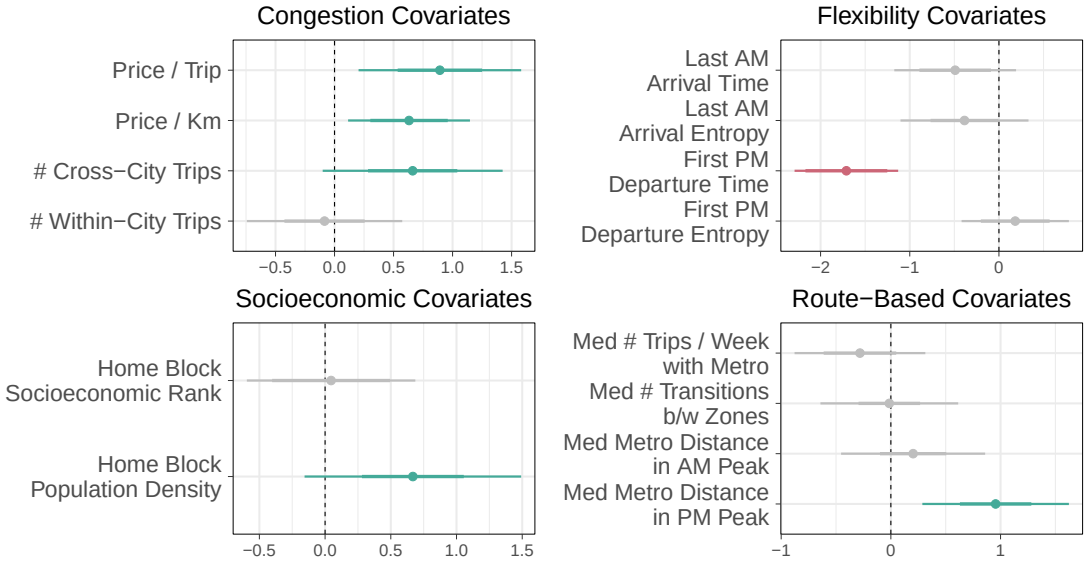


Figure 8: Sorted effects results for treatment effects on the first PM departure time for cross-city trips. The x-axis shows the average difference in covariates, measured in standard-deviation units, between riders with the most positive CATEs (top 20%) and those with the most negative CATEs (bottom 20%) for the outcome.

5 A Model of Time-Varying Traffic Equilibrium

To assess how scaling up the experiment's pricing scheme to the population of potential drivers would affect driver behavior and welfare in equilibrium, we now develop a model of time-varying traffic equilibrium in 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 on both how their different routes intersect and 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.

Overview A mass of Q non-atomic drivers take private vehicle trips through a road network comprised of a graph with nodes $n \in \mathcal{N}$, and a set of directed edges $(n, n') \in \mathcal{E} \subseteq \mathcal{N}^2$ that correspond to road segments connecting these nodes. We track drivers' positions throughout the network in continuous time over a finite interval of length T.

Each road segment e is endowed with a time-varying price function $p_e(\cdot)$, such that a driver who 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 (such as the one used in the experiment) by setting $p_e(\tau)$ to the per-kilometer price of segment e's pricing zone at time τ multiplied by the length of e. 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.

Each driver i is endowed with a single trip (i.e., an origin-destination pair) and a focal time window $\mathcal{T}_i \coloneqq [\underline{T}_i, \overline{T}_i]$ during which they can choose to drive (or not). The identity of the trip and focal time window determine the choice set for how the driver might take their trip. We denote the set of routes available for driver i to take as $\mathcal{J}_i$. Each route $j \in \mathcal{J}_i$ is characterized by a sequence of road segments $r_j \coloneqq (e_{j1}, \ldots, e_{jK_j})$ 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 $e \in \mathcal{E}$.[13]

Driver Choices On a given day, each driver i chooses whether to take their trip, and if so, which route $j \in \mathcal{J}_i$ and departure time $\tau_0 \in \mathcal{T}_i$ to depart on so as to maximize their expected utility $U_{ij}(\tau_0)$:

$$U_{ij}(\tau_0) \coloneqq \underbrace{V_{ij}(\tau_0) - \gamma_i D_j(\tau_0) - \alpha_i P_j(\tau_0)}_{\delta_{ij}(\tau_0)} + \varepsilon_{ij}(\tau_0), \quad U_{i0}(\infty) = \varepsilon_{i0}(\infty).$$

Here, $D_j(\tau_0)$ and $P_j(\tau_0)$ are the expected duration and price of a trip on route j departing at time τ_0. $V_{ij}(\tau_0)$ denotes driver i's "scheduling preference" for leaving at time τ_0 when taking route j. The residual $\varepsilon_{ij}(\tau_0)$ is a Type-I extreme value shock that is i.i.d. across drivers, routes, and times.[14]

Once driver i has departed on route $J_i = j$ at time $T_i = \tau_0$, we model the driver's position along the edges r_j of their route as a time-inhomogeneous, continuous-time Markov chain $E_i(\tau)$. In particular,

[13] This notation assumes that each route traverses each segment in the graph at most once.

[14] For a formalization of type-I extreme value processes for modeling random utility over a continuum of choices, see Dagsvik (1994) and Appendix A of Dupuy and Galichon (2014).

when driver i traverses segment e_{jk} at calendar time τ, they advance to the next route segment $e_{j(k+1)}$ at an instantaneous rate:

$$\lambda_{jk}(\tau) \coloneqq \frac{d}{dh} \Pr\big(E_i(\tau + h) = e_{j(k+1)} \mid E_i(\tau) = e_{jk},\, J_i = j\big)\Big|_{h=0^+} \tag{3}$$

and remain on segment e_{jk} otherwise until they stochastically transition at a later time. The rates $\lambda_{jk}(\tau)$ are determined by contemporaneous congestion conditions, as we will explain below.

A consequence of the Markov chain model is that the amount of time that it takes for driver i to traverse route j is stochastic. That is, while the expected duration of traversing segment e_{jk} starting at time τ is the inverse of the traversal rate $1/\lambda_{jk}(\tau)$, the actual duration for a given traversal is determined by the realization of stochastic transition draws:[15]

$$C_{ik} \coloneqq \inf\{\tau' \geq C_{i(k-1)} \mid E_i(\tau') = e_{j(k+1)}\}.$$

Since prices are charged when a driver exits each segment, we can express the expected trip duration $D_j(\tau_0)$ and price $P_j(\tau_0)$ for a trip departing on route j at time τ_0 as follows:

$$D_j(\tau_0) \coloneqq \mathbb{E}\left[C_{iK_j} - \tau_0 \mid J_i = j,\, T_i = \tau_0\right], \quad P_j(\tau_0) \coloneqq \mathbb{E}\left[\sum_{k=1}^{K_j} p_{e_{jk}}\left(C_{ik}\right) \,\middle|\, J_i = j,\, T_i = \tau_0\right]. \tag{4}$$

Here C_{iK_j} is the time at which driver i completes their trip on route j, so that $D_j(\tau_0)$ is the expected time of the final arrival at i's destination, less the departure time τ_0. $P_j(\tau_0)$ is the expected sum of the segment prices along route j at the stochastic point in time at which each segment is successfully traversed.

The stochastic nature of the model also entails that the progression of a driver's position along their route may violate the rigid "first-in-first-out" physics emphasized by engineering models of traffic flow: it is possible for a driver to traverse a road segment faster than other drivers who entered that same road segment earlier. However, the Markov structure of the route progression process greatly simplifies calculating the level of congestion on each road segment at every point in time: instead of tracking each driver's linear progression along their route—overlapping with different drivers, each on their own route, at different points in space and time—we need only to track the mass of drivers currently located on a given segment in order to know how congested that segment is. This feature allows the model to represent spatiotemporally granular traffic dynamics that are faithful to the physics of traffic flow at the population level across large road networks.

Segment-Level Congestion The canonical model of traffic flow in the transportation engineering literature (see Boyles et al. (2025) for a textbook treatment) assumes that at each instant, the vehicle speed $v_e(\rho)$ (measured in distance per unit of time) on a road segment e is a decreasing function of the vehicle density ρ (measured in vehicles per unit of distance) on that road segment. This yields a relationship often referred to as the *fundamental diagram* of traffic flow, between vehicle density ρ

[15] We define C_{i0} as the departure time T_i.

and traffic flow $f_e(\rho) \coloneqq v_e(\rho) \cdot \rho$ (measured in vehicles per unit of time) through the road segment. Typically, the fundamental diagram $f_e(\rho)$ is concave: flow rises from zero at zero density, peaks at a critical density $\rho_e^{\text{crit}} \coloneqq \arg\max_\rho f_e(\rho)$, and then falls back to zero at a "jam" density ρ_e^{jam}, where traffic is gridlocked and no vehicles can traverse the segment.

Notably, the concavity of the fundamental diagram implies that two different vehicle flow rates can be observed for the same vehicle density and speed. On the increasing branch, the flow is limited by the relative capacity of the road segment: when density is low, a marginal increase in vehicles on the road does not affect speeds very much and so total flow increases more or less proportionally to the increase in the number of vehicles passing through. As density approaches the critical threshold, the road capacity starts to bind: vehicles start to drive closer together and speeds decrease more substantially, leading to a peak in how many vehicles can pass through at a given time. If the road segment were isolated, we would never observe a density higher than the critical threshold, as in queueing-style models (Vickrey, 1969). However, in a network of road segments linked to one another, it is possible for higher densities to arise due to congestion downstream (Small and Chu, 2003).[16] In particular, if the flow rate of a subsequent road segment is too slow, then vehicles traversing the current road segment will be forced to queue up behind the slower-moving vehicles up ahead until the subsequent road segment has enough space to accept them. When the downstream road segment is at capacity, this may cause a spill-back on the current segment that pushes the density past its critical threshold while decreasing the flow rate down to zero at gridlock.

To account for this phenomenon, we take a dynamic approach inspired by the canonical Cell Transmission Model (Daganzo, 1994, 1995; Lebacque, 1996): rather than treating the fundamental diagram as a single function, we split the fundamental diagram at the critical density ρ_e^{crit} into two different flow functions: a *sending flow* function $f_{Se}(\rho)$ that captures the rate at which vehicles exit the road segment and a *receiving flow* function $f_{Re}(\rho)$ that captures the rate at which vehicles can enter the road segment from upstream:

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

Segment-Level Traversal Rates To explain how this segment-level congestion technology determines traversal rates, we denote by $q_{jk}(\tau)$ the mass of drivers who chose route j and 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 $\boldsymbol{q}_j(\tau) \coloneqq \big(q_{j1}(\tau), \ldots, q_{jK_j}(\tau)\big)'$. We define the total mass of drivers $q_e(\tau)$ who are on segment e at time τ and the number of those drivers $q_{ee'}(\tau)$ whose next segment is e' as follows:

$$q_e(\tau) \coloneqq \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e\} q_{jk}(\tau), \quad q_{ee'}(\tau) \coloneqq \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e,\, e_{j(k+1)} = e'\} q_{jk}(\tau). \tag{5}$$

[16] See Anderson and Davis (2020); Hall (2018, 2021) for recent discussions about the empirical prevalence of this phenomenon.

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 upstream and downstream segments could intersect at the same node: the flow from a given upstream segment may enter different downstream segments, and the flow into a given downstream segment may come from multiple upstream segments. We assume that the 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 segment is e'. To capture the spillback effect of downstream congestion, we allocate receiving flow $f_{Re'e}(\boldsymbol{q}(\tau))$ for a downstream segment e' to upstream segment e according to a priority ranking of the upstream segments.[17]

Putting these pieces together, for a movement (e, e') and a time τ, we can write the total flow of drivers transitioning from segment e to the next segment e' on their route as the minimum of the sending flow from e toward e' and the receiving flow that e' grants to upstream segment e:[18]

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

To translate this overall movement flow into the rate $\lambda_{ee'}$ at which any individual driver moves from segment e to e', we divide the total flow $\nu_{ee'}$ by the mass of drivers attempting this transition $q_{ee'}(\tau)$:[19]

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

The driver-level transition rate $\lambda_{jk}(\tau)$ introduced in Equation (3) is given by $\lambda_{ee'}(\boldsymbol{q}(\tau))$ for $e = e_{jk}$ and $e' = e_{j(k+1)}$. Here, the endogenous dependence on the masses of drivers on nearby segments is reflected in the argument $\boldsymbol{q}(\tau)$.

Driver 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 $\theta = (\mathcal{J}, \mathcal{T}, V(\cdot), \alpha, \gamma)$. For a driver of type θ, let $\mathcal{J}(\theta)$ and $\mathcal{T}(\theta) \coloneqq [\underline{T}(\theta), \overline{T}(\theta)]$ denote their route choice set and focal time window, 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 their systematic component of utility. Note that we now make the dependence of expected trip duration $D_j(\tau; \boldsymbol{q})$ and price $P_j(\tau; \boldsymbol{q})$ on $\boldsymbol{q}$ explicit. The aggregate trip departure rate $a_j(\tau; \boldsymbol{q}) \coloneqq Qs_j(\tau; \boldsymbol{q})$

[17] All of the sending flow from an upstream segment 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.

[18] For terminal movements, the flow of drivers completing their trips is given by $\nu_{e:\varnothing}(\boldsymbol{q}(\tau)) \coloneqq f_{Se}(q_e(\tau))\, q_{e\varnothing}(\tau)/q_e(\tau)$. We adopt the convention $0/0 = 0$ for the proportional-allocation terms.

[19] When $q_{ee'}(\tau) = 0$, the derivative of $\nu_{ee'}(\boldsymbol{q}(\tau)) \,/\, q_{ee'}(\tau)$ 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 $\nu_{ee'}(\boldsymbol{q}(\tau))$. We let $\varnothing$ denote the terminal segment for a driver who has completed their trip so $e_{j(K_j+1)} = \varnothing$, and we define $f_{R\varnothing e}(\boldsymbol{q}(\tau)) \coloneqq \infty$ so that receiving flow never acts as a binding constraint for route-terminal movements.

of drivers on route j at time τ is determined by the aggregate route-departure-time choice density $s_j(\tau; \boldsymbol{q})$, which is defined as follows:

$$s_j(\tau; \boldsymbol{q}, \theta) \coloneqq \frac{\exp(\delta_j(\tau; \boldsymbol{q}, \theta))}{1 + \sum_{j' \in \mathcal{J}(\theta)} \int_{\underline{T}(\theta)}^{\overline{T}(\theta)} \exp(\delta_{j'}(u; \boldsymbol{q}, \theta)) du}, \quad s_j(\tau; \boldsymbol{q}) \coloneqq \int_{\Theta} s_j(\tau; \boldsymbol{q}, \theta)\, dF(\theta), \tag{6}$$

where $s_j(\tau; \boldsymbol{q}, \theta)$ denotes the route-departure-time choice density for a particular driver type θ.

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:

$$\begin{aligned} Q_j(\boldsymbol{q}(\tau))[k+1, k] &\coloneqq \lambda_{e_{jk} e_{j(k+1)}}(\boldsymbol{q}(\tau)), \\ Q_j(\boldsymbol{q}(\tau))[k, k] &\coloneqq -Q_j(\boldsymbol{q}(\tau))[k+1, k] \end{aligned} \qquad k \in [K_j],$$

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. Given this structure, the masses of drivers on each segment in the road network propagate along their routes according to the following system of ordinary differential equations (ODEs):

$$\frac{d}{d\tau} \boldsymbol{q}_j(\tau) = Q_j(\boldsymbol{q}(\tau))\, \boldsymbol{q}_j(\tau) + a_j(\tau; \boldsymbol{q})\, \mathbf{e}_1, \quad \boldsymbol{q}_j(0) = \mathbf{0} \quad j \in \mathcal{J}. \tag{7}$$

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 Durations and Prices Using our congestion technology, we now characterize the expected duration $D_j(\tau; \boldsymbol{q})$ and price $P_j(\tau; \boldsymbol{q})$ defined in Equation (4). Let $D_{jk}(\tau; \boldsymbol{q})$ denote the expected remaining trip duration for a driver on route j who is on its k-th segment at time τ. Over a small time increment $d\tau$, a driver on segment k spends $d\tau$ time traveling and either advances to segment $k+1$ with probability $Q_j(\boldsymbol{q}(\tau))[k+1, k] d\tau$ or remains on segment k with the complementary probability.[20] The law of iterated expectations therefore gives

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

Rearranging and taking $d\tau \to 0$ yields the backward ODE

$$-\frac{d}{d\tau} D_{jk}(\tau; \boldsymbol{q}) = 1 + Q_j(\boldsymbol{q}(\tau))[k+1, k] \big(D_{j(k+1)}(\tau; \boldsymbol{q}) - D_{jk}(\tau; \boldsymbol{q})\big), \qquad k \in [K_j]. \tag{8}$$

[20] In a continuous-time Markov chain, the probability of two or more jumps over $[\tau, \tau + d\tau]$ is second order in $d\tau$, so only the zero-jump and one-jump events matter in the first-order expansion.

We solve Equation (8) backward in both time τ, using the horizon-T boundary condition $D_{jk}(T;\boldsymbol{q}) = \bar{D}_{jk}$ for free-flow traversal duration $\bar{D}_{jk}$, and route position k, since the remaining duration of a completed trip is zero: $D_{j(K_j+1)}(\tau;\boldsymbol{q}) = 0$. The expected duration that determines demand in Equation (6) is then given by the continuation value at the first segment: $D_j(\tau;\boldsymbol{q}) = D_{j1}(\tau;\boldsymbol{q})$.

The expected price admits an analogous backward characterization. Let $P_{jk}(\tau;\boldsymbol{q})$ denote the expected price yet to be paid by a driver occupying the k-th segment of route j at time τ. As with duration, the expected price enters a zero absorbing state once a trip is completed: $P_{j(K_j+1)}(\tau;\boldsymbol{q}) = 0$. Now consider that driver's possible position at $\tau + d\tau$. With probability $Q_j(\boldsymbol{q}(\tau))[k+1,k]d\tau$, they complete segment e_{jk}, pay segment price $p_{e_{jk}}(\tau)$, and enter the next segment with continuation price $P_{j(k+1)}(\tau + d\tau;\boldsymbol{q})$. Otherwise, they remain on segment k, make no payment, and carry forward the continuation price $P_{jk}(\tau + d\tau;\boldsymbol{q})$. As such, the law of iterated expectations again gives

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

Collecting terms and taking $d\tau \to 0$ gives

$$-\frac{d}{d\tau}P_{jk}(\tau;\boldsymbol{q}) = Q_j(\boldsymbol{q}(\tau))[k{+}1,k]\big(p_{e_{jk}}(\tau) + P_{j(k+1)}(\tau;\boldsymbol{q}) - P_{jk}(\tau;\boldsymbol{q})\big), \qquad k \in [K_j]. \tag{9}$$

As with duration, we solve Equation (9) backward in both τ and k, using the time-horizon boundary condition $P_{jk}(T;\boldsymbol{q}) = 0$ and the absorbing-state condition $P_{j(K_j+1)}(\tau;\boldsymbol{q}) = 0$. The expected price that enters Equation (6) is the continuation value at the first segment: $P_j(\tau;\boldsymbol{q}) = P_{j1}(\tau;\boldsymbol{q})$.

Equilibrium We define a time-varying traffic equilibrium in our model as a collection of movement-level masses $\boldsymbol{q}^*(\cdot)$ that constitutes a fixed point of an operator $G(\boldsymbol{q}(\cdot))$ that outputs the movement-level masses generated when a candidate $\boldsymbol{q}(\cdot)$ determines congestion and transition rates. The corresponding solutions to Equations (8) and (9) determine expected durations and prices, Equation (6) then determines route departure rates, and the resulting route-level mass propagation given by Equation (7) aggregates back to movement-level masses via Equation (5). In other words, an equilibrium collection of movement-level masses determines congestion and therefore transition rates across segments throughout the network. These transition rates determine the expected durations and prices that drivers face for different departure choices, which in turn determine how likely drivers are to depart at different times and on different routes. The distribution of realized departure choices induces a distribution of driver positions throughout the network over time, which aggregates up to the equilibrium movement-level masses.

Welfare Metrics Given an equilibrium $\boldsymbol{q}^*(\cdot)$ and a revenue recycling parameter Λ, we evaluate average consumer surplus CS, revenue R, and money-metric welfare W_Λ as follows (Train, 2009):

$$\begin{aligned} CS &:= \int_\Theta \underbrace{\frac{1}{\alpha} \log \left[1 + \sum_{j \in \mathcal{J}(\theta)} \int_{\underline{T}(\theta)}^{\overline{T}(\theta)} \exp\big(\delta_j(\tau; \boldsymbol{q}^*, \theta)\big)\, d\tau \right]}_{CS(\theta)} dF(\theta), \\ R &:= \sum_{j=1}^{J} \int_0^T s_j(\tau; \boldsymbol{q}^*) P_j(\tau; \boldsymbol{q}^*)\, d\tau, \qquad W_\Lambda := CS + \Lambda R, \end{aligned} \tag{10}$$

where $CS(\theta)$ denotes the expected consumer surplus of a driver of type θ.

Implementation Although our model is written in continuous time, computing equilibria in practice requires discretization. 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 $\lambda_{ee'}(\bar{\boldsymbol{q}}_t)$ when evaluated at the period-average movement-level masses $\bar{q}_{ee't}$:

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

We approximate each driver's continuous choice set by a finite set of route-and-departure-period alternatives under the assumption that utility is constant and departure times are uniform within each departure period. This yields the mixed logit demand system over discrete departure windows that we estimate in Section 6.1. Under these approximations, the ODE systems governing mass propagation in Equation (7) and expected trip durations and prices in Equations (8) and (9) can be expressed as systems of linear equations. The period-average movement-level masses $\bar{q}_{ee't}$ can also be expressed as linear functions of start-of-period route-level masses. These linear relationships yield 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.[21] See Appendix Section G.1 for details.

While the linear systems of equations that govern the discretized equilibrium dynamics do not admit closed forms, we describe how to evaluate them in a computationally efficient manner using a technique called uniformization (Stewart, 1994) in Appendix Section G.2. We provide a complete description of the accelerated fixed point algorithm that we use to solve for discretized equilibria in Appendix Section G.3. In Appendix Section G.4, we describe the optimization algorithm that we use to maximize welfare W_Λ with respect to prices under different levels of revenue recycling and different types of pricing rules. In Appendix Section G.5, we describe the procedure that we use to compute the smallest revenue recycling parameter Λ for which non-zero congestion prices are welfare-improving.

[21] Unlike other traffic equilibrium models that require space and time discretizations to be carefully aligned for numerical stability (see Boyles et al. (2025)), our model allows the granularity of space and time discretizations to be chosen independently depending on which is more important in a particular application.

6 Model Estimation

In order to take our model to the data, we first estimate the distribution of primitives that underpin supply and demand for driving across the road network. For demand, we estimate the preferences over prices and durations that govern the decision to drive for different individuals in our experimental panel. For supply, we estimate the relationship between traffic density and flow on specific road segments. We then connect demand and supply using a population-weighted mapping of driving through highways by our panelists to the concurrent traffic flows observed on those highways overall.

6.1 Demand for Trips

In order to identify preferences over driving conditions, 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 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. See Appendix Section D.1 for details.

In order to allow preferences to vary flexibly across travel purposes and times of day, we refer to each habitual 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 $\{(j,t) : j \in \mathcal{J}_i, t \in \mathcal{T}_i\}$ 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.[22] 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 and time in our panel using high frequency GPS pings from the Israeli public bus network, as detailed in Appendix Section D.2. 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. The expected duration $D_{i\ell t}$ for each trip

[22] 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.

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.

The imputed road speeds also determine the expected prices $P_{i\ell t}$. For each trip opportunity (i, ℓ, t), we sum the prices charged for each kilometer in trip i's route depending on which time bin the kilometer is predicted to be driven in. If slower road speeds at time t entail that a kilometer in the metro zone is driven during peak hours instead of off-peak, for instance, then we account for the higher resulting price in $P_{i\ell t}$. We also add the expected cost of gasoline based on the number of kilometers in i's route under average fuel efficiency and time varying gasoline prices. When drivers are in "monitoring" status, the price $P_{i\ell t}$ only includes the gasoline cost. When drivers are in "active" status, the price $P_{i\ell t}$ includes both gasoline and the experiment's congestion prices. See Appendix Section D.3 for details

In order to account for heterogeneity in the availability of attractive public transit alternatives, we assign each habitual trip i a public transit ratio R_i that is analogous to the public transit index used in Section 4.2. We compute R_i as the ratio of the projected driving duration to the duration of the best available public transit alternative at the median departure time in the driver's choice set; see Appendix Section D.5. Because R_i is evaluated at the fixed median time rather than recomputed for each potential departure time in the choice set, it is a trip-level characteristic rather than a choice-specific characteristic like expected trip duration or price. Because public transit duration enters the denominator, a higher ratio indicates a more attractive public transit alternative. Notably, only 1.1% of trips have a viable public transit alternative involving rail (without first/last leg driving) and nearly 40% of trips require at least one bus transfer; see Appendix Table A4 for a detailed breakdown of best transit modes.

Our panel of trip opportunities across drivers allows us to estimate mixed logit preferences at a granular level as in 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, price, and transit ratio coefficients $(\gamma_i, \alpha_i, \rho_i)$ as a joint Normal distribution with a mean vector depending linearly on a vector of demographic variables X_i (in particular, standardized home-neighborhood population density and home-neighborhood socio-economic status as described in Section 3) and a general covariance matrix Σ. We estimate the parameters of our demand model by minimizing a panel mixed logit negative log-likelihood objective with an added IV-GMM "penalty" akin to the CLEER estimator proposed by Grieco et al. (2025); see Appendix Section E for details.

The identification for our parameter distributions relies on both panel variation—including the within-driver experimental status change from "monitoring" to "active"—and cross sectional variation in exposure to coarse pricing bins, higher gas prices, and the availability of public transit alternatives. Importantly, market-mean duration could plausibly be positively correlated with drivers' unobserved scheduling preferences, since times when more drivers want to depart (e.g., rush hour tends to be at the time when typical workdays end) are also the times when trip durations are the longest due to congestion. To address this endogeneity concern, we construct an instrument that takes advantage of spatial dispersion of sporadic rainfall events. As documented by the press (The Times of Israel, 2020, 2021), 2020 and 2021 had unusually high rainfall throughout the calendar year—often separating nearby regions with dramatically different rain exposure. Because heavy precipitation is

relatively rare in Israel, major rainstorms have a substantial impact on road conditions and generally slow down driving speeds. To leverage the variation from this supply shock across different trip opportunities, we construct an instrument for trip duration using the distance-weighted average of cumulative precipitation across the TAZs traversed by trip i. See Appendix Section D.4 for details.

We report the results of our demand estimation procedure in Table 1. On average, drivers dislike both higher prices (α_0) and longer travel times (γ_0), with both mean taste coefficients estimated precisely. These sensitivities vary systematically with driver demographics: drivers from higher socio-economic-status neighborhoods are less price-sensitive but more duration-sensitive, while drivers from denser neighborhoods—where public transit tends to be more accessible—are both less price-sensitive and less duration-sensitive. Having better transit alternatives has a noisy effect on average, but it reduces driving more for drivers with higher home-neighborhood socio-economic status and population density. Beyond this demographic variation, every entry of the estimated random-coefficient covariance matrix Σ is statistically distinguishable from zero, indicating substantial remaining heterogeneity in preferences that is unexplained by observed demographics. To benchmark our estimates, we find that the median Value of Time γ_i/α_i (measured in NIS per hour) implied by our estimates is 50 NIS, which is close to a reported estimate of the median hourly wage for hourly workers (Dagan-Buzaglo, 2025) or 150% of the minimum hourly wage (National Insurance Institute of Israel, 2026) and well within typical relative-to-hourly-wage estimates found in the literature.[23] See Appendix Section A.3 for the full distribution.

6.2 Supply of Traffic Flow

Taking the supply side of the model in Section 5 to the data requires three sets of inputs. We first estimate the segment-level congestion functions f_e that map the density of vehicles on road segment e into the rate at which vehicles traverse it. We then construct the road network $(\mathcal{N}, \mathcal{E})$ over which these flows interact. Finally, we connect supply and demand by translating the habitual trips in our experimental panel into the total mass of potential drivers Q and the distribution of driver types F.

Estimating Segment-Level Congestion Technologies We estimate the congestion functions f_e using observed lane-level traffic volumes (flows) and median speeds in 5-minute intervals from the Ayalon roadway sensors described in Section 3.2. We use weekday data from 2020 through 2022, yielding approximately 1.2 million observations from 24 sensors. For each sensor-lane-interval observation, we recover traffic density by dividing the observed flow by the median speed, in keeping with the discussion of the relationship between traffic flow, density, and speed in Section 5.

Because the available sensor data covers only a subset of the roads used in our counterfactuals, we estimate per-lane congestion functions for classes of observably similar roads. We classify sensors by their speed limits (70 or 90 kilometers per hour) and whether they are located on a regular roadway or a bridge or tunnel, producing four baseline road classes. Within each class, we estimate the lane-level flow-density relationship nonparametrically using a constrained least squares procedure. In particular, we model each relationship as a flexible Bernstein polynomial subject to three linear shape restrictions

[23] 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)	-0.1476^{***}	(0.0015)	$[-0.1506, -0.1446]$
Duration (γ_0)	-4.9773^{***}	(0.3869)	$[-5.7356, -4.2191]$
Transit Ratio (ρ_0)	-0.2442	(0.2077)	$[-0.6514, 0.1629]$
Interactions with Price			
SES × Price	0.0415^{***}	(0.0009)	[0.0398, 0.0432]
Population Density × Price	0.0115^{***}	(0.0007)	[0.0100, 0.0130]
Interactions with Duration			
SES × Duration	-1.0920^{***}	(0.0313)	$[-1.1533, -1.0307]$
Population Density × Duration	0.1137^{***}	(0.0298)	[0.0553, 0.1721]
Interactions with Transit Ratio			
SES × Transit Ratio	-0.3595^{***}	(0.0214)	$[-0.4014, -0.3176]$
Population Density × Transit Ratio	-0.1789^{***}	(0.0171)	$[-0.2123, -0.1454]$
Random Coefficient Covariance (Σ)			
Var(Price) ($\Sigma_{p,p}$)	0.0022^{***}	(0.0001)	[0.0019, 0.0025]
Cov(Duration, Price) ($\Sigma_{d,p}$)	-0.0405^{***}	(0.0046)	$[-0.0496, -0.0314]$
Cov(Transit Ratio, Price) ($\Sigma_{\text{tr},p}$)	0.0165^{***}	(0.0035)	[0.0097, 0.0234]
Var(Duration) ($\Sigma_{d,d}$)	1.6347^{***}	(0.0754)	[1.4869, 1.7826]
Cov(Transit Ratio, Duration) ($\Sigma_{\text{tr},d}$)	0.2578^{***}	(0.0760)	[0.1087, 0.4068]
Var(Transit Ratio) ($\Sigma_{\text{tr,tr}}$)	0.5277^{***}	(0.1067)	[0.3186, 0.7367]

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

Table 1: Panel mixed logit demand estimates. Standard errors are clustered by habitual trip and market; see Appendix Section E.3 for details.

based on the properties of the fundamental diagram: flow is concave in density, flow is zero when density is zero, and the slope at zero density equals a free-flow speed proxy based on the class's speed limit. Figure 9 plots an estimated function for one of the four road classes.

We extend these estimates to the remaining major roads we consider in our counterfactuals using the road segments' characteristics. For speed limits between 50 and 120 kilometers per hour, we linearly interpolate or extrapolate between the functions estimated for the 70 and 90 kilometer-per-hour classes. For a segment e with L_e lanes and total density ρ, we then aggregate the lane-level congestion function f_e^{lane} to the segment level by scaling its input and output:

$$f_e(\rho; L_e) = L_e \, f_e^{\text{lane}}\left(\frac{\rho}{L_e}\right). \tag{11}$$

See Appendix Section F.1 for details.

Constructing the Counterfactual Road Network Next, we construct the road network used in our counterfactuals from the OSM data introduced in Section 3.2. We allow endogenous congestion on all major roads in the largest commuting zone in Israel, the "Central District" around Tel Aviv (visualized in Figure F3) and use the 17,748 habitual trips between February and July of 2021 whose imputed routes use at least one of these major roads and whose choice sets lay between 2 PM and 10 PM. Although we only allow for endogenous congestion on major roads, we account for the prices and

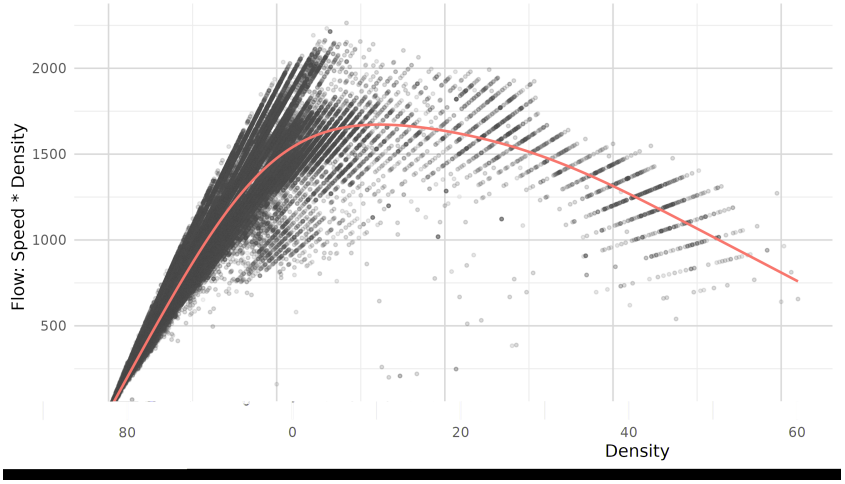


Figure 9: An estimated congestion function relating traffic flow on a segment of the Ayalon highway to traffic density on that segment.

durations incurred on all other OSM road segments traversed by habitual trips by assuming that these segments' per-vehicle traversal rates are fixed to those implied by the imputed segment-level speeds that we use to determine potential trip durations in Section 6.1. The resulting road network tracks endogenous vehicle masses on 4,707 edges. See Appendix Section F.2 for details.

Connecting Trip Demand to Traffic-Flow Supply We use two external data sources to make the focal habitual trips representative of aggregate traffic. Cell-phone-data-based origin-destination trip counts by hour, described in Section 3.2, determine the composition of trips across locations and times of day. Highway sensor volumes then determine the absolute traffic scale relevant for the estimated congestion functions.

We first choose nonnegative population weights w_i for the focal habitual trips through the Central District so that their weighted average weekday realizations reproduce the cell-phone-data-based number of car trips in each supported origin-destination-hour cell. These restrictions determine the traffic represented by the panel in each cell but generally do not determine how that traffic should be allocated across the individual habitual trips that can contribute to it. We resolve this indeterminacy using a variant of entropy balancing (Jaynes, 1957): among the weights that satisfy the origin-destination-hour restrictions, we select those closest to a reference allocation that distributes each home area's adult population equally across the supported habitual trips of drivers who live there. This procedure preserves the population composition of drivers' home locations to the extent permitted by the external trip-flow restrictions.

The origin-destination weights determine relative traffic demand across space and time, but their scale may not reflect the vehicle-flow scale of the estimated congestion functions. To reconcile the two, we map each Ayalon sensor to a directed road-network edge and use the inferred routes for habitual trips to determine which realized trips in our data would cross it. We then aggregate both predicted crossings and observed sensor volumes to 30-minute intervals and regress observed sensor volumes on the weighted number of predicted crossings without an intercept. The resulting common coefficient

The speed gains generated by the experiment fee schedule are not fully driven by the extensive margin. To see this, we consider a stylized version of the model in which drivers' trip probabilities are fixed so that their only choice is when to take their trip. With only time-shifting as an option to reduce their exposure to pricing, drivers pay slightly more per trip and the increase in rebated consumer surplus is halved, but average trip speeds increase by 3.75 km/h—two-thirds of the gain under the full model.[26]

Benchmarking to Road Expansion Although our paper is primarily focused on road pricing, our model allows us to evaluate other types of interventions that might reduce congestion and increase speeds and welfare. As a benchmark, we consider a counterfactual in which no pricing is imposed, but every highway in the Central District is expanded with a new lane in each direction. Table 3 reports the corresponding equilibrium effects. On average, adding a lane increases the probability of driving by about 4 percentage points (12% of the baseline probability) and average trip speeds by 3.92 km/h (9% of the baseline). Because no tolls are collected, revenues are irrelevant but consumer surplus increases by 1.73 NIS per trip-day due to the speed gains and added trips. That is, without taking into account the costs of land provision and road construction, congestion prices based on the experiment fee schedule with full revenue recycling would generate a net welfare gain equivalent to 93% of the gain from adding a lane to every single highway.

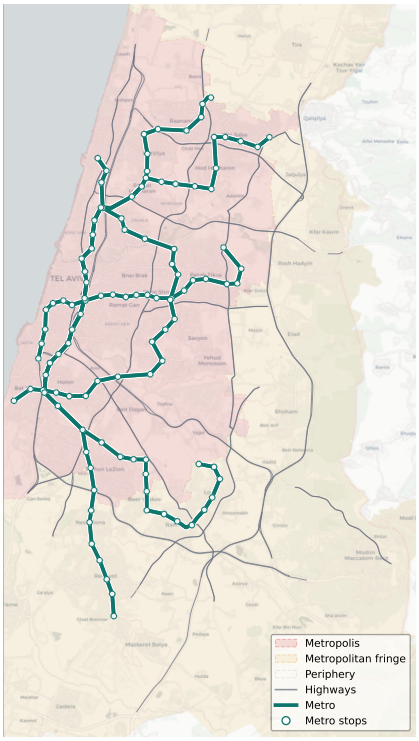
	Mean	SD	Q25	Median	Q75
Trip Probability (%)	3.93	5.62	0.06	1.54	5.58
Distance / Day (km)	0.89	2.47	0.00	0.14	0.75
Price / Trip (NIS)	0.00	0.00	0.00	0.00	0.00
Duration / Trip (mins)	-2.46	3.25	-3.76	-1.14	-0.04
Speed / Trip (km/h)	3.92	4.25	0.24	2.81	6.45
Raw Consumer Surplus (NIS/day)	1.73	5.79	0.01	0.28	1.42
Rebated Consumer Surplus (NIS/day)	1.73	5.79	0.01	0.28	1.42

Table 3: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium after adding one highway lane, relative to the baseline equilibrium with no intervention.

Benchmarking to Transit Expansion Another commonly cited intervention for reducing congestion is to expand transit infrastructure. To simulate a benchmark in this direction, we consider the planned Tel Aviv metro (subway) system, which began construction in 2025 and is set to open in 2037 (Wikipedia, 2026). Table 4 summarizes the effects of the new metro system across the Central District. Only 8% of trips in the Central District would have a faster transit route using the metro system—2% with a direct route and 6% with a route that combines the metro with bus or rail. Among these, the metro system would reduce transit time by about 5 minutes on average, increasing the driving-to-transit ratio by 12-16%. Despite these gains, our model predicts that equilibrium driving probabilities would respond modestly. Drivers with a direct metro alternative reduce their driving probability by

[26] See Appendix Tables A7 and A8 for a trip-level summary and segment-level speed changes by baseline occupancy.

Table 4: Metro expansion: planned network and driver outcomes by metro route type. The left panel maps the planned metro lines across the Central District. The right panel reports baseline driving cost and time, transit-time savings from the new metro, and equilibrium changes relative to the no-price baseline, for drivers grouped by whether their fastest transit itinerary would use the metro alone (direct), in combination with bus or rail (combo), or not at all (not improved).



	Population Share (%)	Mean	SD	P25	P50	P75
Baseline Avg. Driving Cost (NIS)						
Direct metro (subway + walk)	2.2	5.06	2.83	3.29	4.07	6.56
Combo metro (subway + bus/rail)	6.0	7.61	4.64	4.87	6.80	9.15
Not improved by metro	91.9	7.60	9.79	2.61	4.95	8.61
Baseline Avg. Driving Time (mins)						
Direct metro (subway + walk)	2.2	31.00	14.33	20.15	26.93	38.35
Combo metro (subway + bus/rail)	6.0	36.23	15.77	24.33	33.92	45.74
Not improved by metro	91.9	31.56	23.50	17.26	25.98	37.71
Transit Mins Saved with Metro						
Direct metro (subway + walk)	2.2	4.76	5.36	0.28	3.10	6.92
Combo metro (subway + bus/rail)	6.0	4.89	4.45	1.52	3.80	6.83
Not improved by metro	91.9	0.00	0.00	0.00	0.00	0.00
% Driving-to-Transit Ratio Increase with Metro						
Direct metro (subway + walk)	2.2	16.02	19.47	0.96	8.31	23.25
Combo metro (subway + bus/rail)	6.0	12.02	12.02	3.35	8.30	16.91
Not improved by metro	91.9	0.00	0.00	0.00	0.00	0.00
Δ Eqm. Trip Probability (pp)						
Direct metro (subway + walk)	2.2	-1.24	3.16	-1.62	-0.18	-0.00
Combo metro (subway + bus/rail)	6.0	-0.26	1.63	-0.63	-0.13	0.07
Not improved by metro	91.9	-0.00	0.41	-0.00	0.00	0.01
Δ Eqm. Speed / Trip (km/h)						
Direct metro (subway + walk)	2.2	0.10	0.24	0.00	0.01	0.12
Combo metro (subway + bus/rail)	6.0	0.04	0.18	-0.01	0.00	0.04
Not improved by metro	91.9	0.01	0.12	-0.00	0.00	0.02
Δ Eqm. Consumer Surplus (NIS/day)						
Direct metro (subway + walk)	2.2	0.91	3.45	0.02	0.19	0.81
Combo metro (subway + bus/rail)	6.0	0.17	1.64	-0.01	0.09	0.37
Not improved by metro	91.9	0.02	0.35	-0.00	0.00	0.00

1.24 percentage points, while drivers with a combo alternative reduce their driving probability by only 0.26 percentage points. These drivers do benefit from the metro system: consumer surplus increases by 0.91 NIS per trip-day for direct metro drivers and 0.17 NIS per trip-day for combo drivers—roughly half and one tenth of the average rebated consumer surplus from congestion pricing with the experiment fee schedule. However, since these trips represent a small minority of drivers, the total increase in consumer surplus is only 0.04 NIS per trip-day.

Optimal Pricing with Revenue Recycling A natural question is how much better congestion pricing might do if the prices were set optimally. To answer this question, 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 under full revenue recycling ($\Lambda = 1$) using the price optimization algorithm described in Appendix Section G.4.[27] The left panel of Table 5 reports the resulting optimal usage-based price schedule. The optimal price for kilometers driven in the metro zone during peak hours is close to the price set by the experiment: 1.69 NIS / km relative to the experiment's 1.5 NIS / km. However, while metro fringe peak is slightly lower than in the experiment

[27]It would be possible, in principle, to find optimal prices for every road segment and time interval. However, while we are comfortable ignoring routing choices under the coarse geographic zones used in the experiment, it is likely that more granular pricing would incentivize substantial rerouting. Since we don't observe direct routing choices in our data, and don't account for routing choice in the demand model, allowing for rerouting would require a substantial extension of our current demand estimation routine.

(a) Usage-based

	Peak	Off-peak
Metropolis	1.69	0.99
Metro Fringe	0.23	0.00
Periphery	2.28	0.03

(b) Cordon-based

	Peak	Off-peak
Metropolis → Metro Fringe	9.33	9.36
Metro Fringe → Metropolis	10.78	13.77
Metro Fringe → Periphery	2.17	0.86
Periphery → Metro Fringe	0.11	0.18

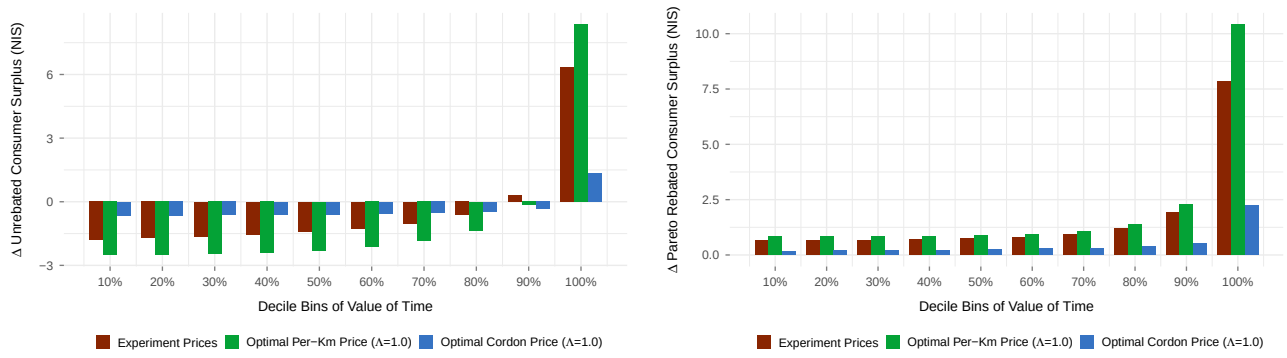
Table 5: Optimal price levels with full revenue recycling ($\Lambda = 1$) under usage-based and cordon-based pricing.

price schedule, metro off-peak and periphery peak are both substantially higher. As a result, average driving probability decreases by nearly 9 percentage points relative to the equilibrium without pricing, and average rebated welfare increases by 2 NIS per trip-day—a 17% improvement over the welfare from adding a lane to every highway in the Central District. See Appendix Table A8 for a summary of the equilibrium effects on driver-level trip outcomes and money-metric welfare.

Optimal Cordon Pricing with Revenue Recycling 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. 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. Unsurprisingly, the largest tolls are at crossings to and from the metro zone, which is also priced highest in the usage based scheme. More surprisingly, the tolls are higher for crossings into metro and in the off-peak hours than from the metro. This reflects the different nature of the trips being targeted by the directional crossing: trips exiting the metro in the evening are likely returning commuting trips with lower elasticities of demand. On the whole, however, cordon prices are much less effective at improving welfare than usage-based prices. Average trip speeds increase by under 2 km/h, compared to 8 km/h under usage-based prices, and average rebated welfare increases by 0.48 NIS per trip-day. See Appendix Table A9 for a summary of the equilibrium effects on driver-level trip outcomes and money-metric welfare.

Distributional Effects and Pareto Improvements In all three of the pricing scenarios considered above, the average raw consumer surplus impact is negative. This suggests that revenues play an important role in generating the welfare gains from pricing. Figure 11 makes this clear by showing the distribution of per-driver welfare changes across value-of-time decile bins with and without revenue recycling. Without revenue recycling, nearly all drivers are worse off, but the top decile of drivers with the highest values of time gain disproportionately more. This pattern is most extreme in the optimal usage-based pricing scheme, in which 80% of drivers are worse off relative to the baseline without pricing in terms of raw consumer surplus. But the optimal usage based pricing scheme also generates the highest revenues—2.96 NIS per trip-day—which is more than enough to compensate the losses and generate a Pareto improvement.

Panel (b) shows the corresponding distribution of welfare changes when revenues are rebated so



(a) No revenue recycling

(b) Pareto-improving revenue recycling

Figure 11: Distribution of per-driver welfare changes by value-of-time decile under different pricing schemes.

as to compensate each driver for their consumer surplus loss relative to the no-price baseline and then distribute the remainder equally among all drivers. Although optimal usage-based prices generate the biggest losses in raw consumer surplus, they also generate the highest rebated welfare for every decile of value of time. By contrast, optimal cordon prices generate substantially fewer losses—only 35% of drivers need to be compensated for a Pareto improvement—but also generate substantially less revenue and ultimately lower rebated welfare for every bin of drivers.[28] This reflects the limited targeting ability of cordon pricing: it cannot distinguish drivers briefly passing through highly congested zones from heavy users of congested roads and so it must charge lower average prices overall.

Appendix Figure A9 plots the distribution of raw and rebated welfare changes by decile of socioeconomic status. As with value of time, raw consumer surplus decreases on average for most SES bins, at the benefit of richer drivers (albeit less evenly and with the highest benefits accruing to the 80-90th percentile of drivers). However, the Pareto improving distribution of rebated welfare is evenly distributed across the top 70% of SES bins and proportionally higher for the bottom 30% as well.

Optimal Prices without Revenue Recycling As the discussion of optimal prices above suggests, revenues play an important role in efficiently reducing congestion. However, as we showed in our motivating example in Section 2, they are not always necessary to generate a net welfare improvement from congestion pricing. As Appendix Table A12 shows, it is possible to improve net welfare with a usage based pricing schedule even when no revenues are included in welfare at all. The optimal price is positive for just one category: periphery-peak at 2.34 NIS / km. This suggests that, as in our example in Figure 1, most trips that involve the periphery also touch the much more congested metro and metro fringe. Charging for periphery driving during peak hours therefore targets long trips that cross multiple zones but are identifiable by their traversals of the periphery. Because these trips are also expensive to begin with due to gas costs, and their drivers appear to be relatively elastic, targeting them and not the drivers who stay within the metro or metro fringe can efficiently reduce congestion in those zones, modestly improving net welfare.

[28] See Appendix Table A10 for a breakdown of Pareto-improving revenue recycling under each pricing scheme.

Table 6: Smallest revenue-share weight Λ^* such that positive prices are optimal.

	Full network het.		**No network het.**
Preferences	Usage-based	Cordon	
Estimated	0.00	0.00	0.21
Homogeneous	0.44	0.59	0.57

The Role of Heterogeneity As the motivating example in Section 2 highlights, the gains from this sort of targeting can only be realized if drivers are sufficiently heterogeneous—either in preferences, in exposure to congested roads, or both. In Table 6 we show that this holds in our setting as well. To do so, we simulate counterfactual networks with each of the two types of heterogeneity shut down: homogeneous preferences keep drivers' routes and choice sets fixed, but use identical demand parameters set at the population average for everyone; no network heterogeneity simulates drivers who face ex-ante random route and departure time sets; see Appendix Section G.6 for details. As the table shows, it is possible to generate net welfare improvements with no revenue recycling with both usage based and cordon pricing under the baseline network.[29] However, if drivers all had the same preferences, then at least 44% of revenues would need to be recycled in order to generate welfare improvements under usage based pricing, and 59% under Cordon pricing. With no network heterogeneity, both numbers are higher: if preferences are at baseline, then at least 21% of revenues would be needed to generate welfare improvements; if preferences are homogeneous as well then at least 57% of revenues would be needed. This logic extends beyond the margin as well: optimal usage based prices with revenue recycling without preference heterogeneity can only generate an average welfare gain of 0.71 NIS per trip-day—35% of the gains under full heterogeneity. If all heterogeneity is shut down, then optimal pricing can only generate an average welfare gain of 0.10 NIS per trip-day.[30]

8 Conclusion

Congestion pricing is among the most heavily studied and most heavily resisted tools in transportation policy. The core of the resistance is an intuitive concern: that added tolls would induce heavy losses to drivers but deliver only modest benefits through time savings. In this paper, we ask what conditions are necessary for a congestion pricing policy to be net beneficial. Our theoretical framework makes clear that congested roads alone are not enough to justify congestion pricing: tolls either need to be recycled back to consumers, or they need to efficiently sort drivers with different values for driving under congested conditions. When one or both of these forces is at play, the nature and value of optimal congestion pricing depends on the primitives of supply and demand: how elastic different drivers are to price and trip duration; how elastic road speeds are to congestion; and how much of the revenues collected from tolls can be recycled back to consumers.

To make this concrete, we assess the feasibility of different congestion pricing policies in the Central

[29] Note that the optimal Cordon prices when $\Lambda = 0$ are only marginally above zero.

[30] See Appendix Table A14 for the optimal pricing schedules under homogeneous demand, and Appendix Tables A15–A17 for the corresponding equilibrium effects under homogeneous demand and under no network heterogeneity.

District of Israel. Using data from a large usage-based pricing experiment, we show that Israeli drivers are elastic to price, but systematically heterogeneous in their preferences and margins of subsitution. If the experiment's fee schedule were to become policy, the number of trips taken in the Central District would decrease by 15% and trip speeds would increase by 13%. To make this equilibrium a welfare improvement, at least some of the toll revenues would need to be recycled back to consumers. With full recycling, however, the policy could generate a Pareto improvement, making every potential driver better off.

Our model allows us to consider alternative transit interventions and to benchmark the welfare gains from congestion pricing against them. We find, for instance, that the experiment's fee schedule would generate 93% of the gains from adding a lane to every highway in the Central District and substantially more than the gains from the Tel Aviv metro currently under construction. Optimal usage-based prices would charge higher fees in the periphery during peak hours and inside the metro during off-peak hours, delivering higher net welfare gains under revenue recycling. It is possible to generate welfare improving congestion prices even without revenue recycling—by charging kilometers driven in the periphery as a way to target a sliver of drivers who also drive in the more congested metro and metro fringe zones. Pricing Cordon gate crossings instead of every kilometer driven makes it more difficult to induce efficient sorting and so the net welfare benefit is lower than with usage-based pricing. That said, either approach can generate welfare gains with or without revenue recycling. As our framework suggests, the latter result is attributable to the extent of heterogeneity in preferences and network exposure across drivers in our setting. Without preference heterogeneity, at least 44% or 59% of revenues would need be recycled under usage-based and Cordon pricing, respectively, in order to make any congestion pricing policy net welfare improving. Without heterogeneity in network exposure either, the minimum revenue recycling rate would be 57%.

References

Akbar, P. A., V. Couture, G. Duranton, and A. Storeygard (2023): "The Fast, the Slow, and the Congested: Urban Transportation in Rich and Poor Countries," Working Paper 31642, National Bureau of Economic Research.

Akbar, P. A. and G. Duranton (2017): "Measuring the Cost of Congestion in a Highly Congested City: Bogotá," Working Paper 2017/04, CAF–Development Bank of Latin America.

Allen, T. and C. Arkolakis (2022): "The Welfare Effects of Transportation Infrastructure Improvements," *The Review of Economic Studies*, 89, 2911–2957.

Allen, T., S. Fuchs, and W. F. Wong (2025): "Evaluating Transportation Improvements Quantitatively: A Primer," Working Paper 2025-13, Federal Reserve Bank of Atlanta.

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.

Anderson, M. L. (2014): "Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion," *The American Economic Review*, 104, 2763–2796.

Anderson, M. L. and L. W. Davis (2020): "An empirical test of hypercongestion in highway bottlenecks," *Journal of Public Economics*, 187, 104197.

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.

——— (1994): "The welfare effects of congestion tolls with heterogeneous commuters," *Journal of Transport Economics and Policy (JTEP)*, 28, 139–161.

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized Random Forests," *The Annals of Statistics*, 47, 1148–1178.

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.

Billings, S. B. (2011): "Estimating the Value of a New Transit Option," *Regional Science and Urban Economics*, 41, 525–536.

Bordeu, O. (2025): "Commuting Infrastructure in Fragmented Cities," Working paper.

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting Event-Study Designs: Robust and Efficient Estimation," *The Review of Economic Studies*, 91, 3253–3285.

Boyd, S. and L. Vandenberghe (2004): *Convex Optimization*, Cambridge University Press.

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

Brancaccio, G., M. Kalouptsidi, and T. Papageorgiou (2026): "Investment in Infrastructure and Trade: The Case of Ports," *Journal of Political Economy*, forthcoming; NBER Working Paper No. 32503.

Brinkman, J. C. (2016): "Congestion, Agglomeration, and the Structure of Cities," *Journal of Urban Economics*, 94, 13–31.

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*.

Chen, J., D. L. Chen, and G. Lewis (2021): "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models," Working Paper.

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

Cohen, Y., R. Maio, and Y. Katz (2021): "Trip Volume and Distribution in Israel, 2018–2019: Final Report," Report 4, Israel Ministry of Transportation and Road Safety, national Model for Travel Demand; prepared by Matat – Transportation Planning Center Ltd., in collaboration with Kolnick and Mobi.

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," Working paper.

Couture, V., G. Duranton, and M. A. Turner (2018): "Speed," *Review of Economics and Statistics*, 100, 725–739.

Csiszár, I. (1975): "I-divergence geometry of probability distributions and minimization problems," *The annals of probability*, 146–158.

Dagan-Buzaglo, N. (2025): "Employment by the Hour is Harmful Employment," .

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.

Dagsvik, J. K. (1994): "Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes," *Econometrica: Journal of the Econometric Society*, 1179–1205.

Deville, J.-C. and C.-E. Särndal (1992): "Calibration estimators in survey sampling," *Journal of the American statistical Association*, 87, 376–382.

Deville, J.-C., C.-E. Särndal, and O. Sautory (1993): "Generalized raking procedures in survey sampling," *Journal of the American statistical Association*, 88, 1013–1020.

Dipasquale, D., E. L. Glaeser, A. M. Guren, and P. S. Willen (2026): "Prices Versus Quantities Revisited: What Do Policymakers Need to Know to Set Pigouvian Taxes and Subsidies?" Tech. rep., National Bureau of Economic Research.

Donaldson, D. (2018): "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," *American Economic Review*, 108, 899–934.

——— (2025): "Transport Infrastructure and Policy Evaluation," in *Handbook of Regional and Urban Economics*, ed. by D. Donaldson and S. J. Redding, Elsevier, vol. 6, 287–352.

Dupuy, A. and A. Galichon (2014): "Personality traits and the marriage market," *Journal of Political Economy*, 122, 1271–1319.

Duranton, G. and M. A. Turner (2011): "The fundamental law of road congestion: Evidence from US cities," *American Economic Review*, 101, 2616–2652.

Durrmeyer, I. and N. Martinez (2022): "The welfare consequences of urban traffic regulations," Working paper, Toulouse School of Economics.

Fajgelbaum, P. D. and E. Schaal (2020): "Optimal Transport Networks in Spatial Equilibrium," *Econometrica*, 88, 1411–1452.

Fogel, R. W. (1964): *Railroads and American Economic Growth: Essays in Econometric History*, Baltimore: Johns Hopkins Press.

Fosgerau, M. and A. de Palma (2013): "The Dynamics of Urban Traffic Congestion and the Price of Parking," *Journal of Public Economics*, 105, 106–115.

Graham, B. S., C. C. d. X. Pinto, and D. Egel (2012): "Inverse probability tilting for moment condition models with missing data," *The Review of Economic Studies*, 79, 1053–1079.

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

Hainmueller, J. (2012): "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political analysis*, 20, 25–46.

Hall, J. D. (2018): "Pareto improvements from Lexus Lanes: The effects of pricing a portion of the lanes on congested highways," *Journal of Public Economics*, 158, 113–125.

——— (2021): "Can tolling help everyone? Estimating the aggregate and distributional consequences of congestion pricing," *Journal of the European Economic Association*, 19, 441–474.

Henderson, J. V. (1974): "Road Congestion: A Reconsideration of Pricing Theory," *Journal of Urban Economics*, 1, 346–365.

Herzog, I. (2024): "The City-Wide Effects of Tolling Downtown Drivers: Evidence from London's Congestion Charge," *Journal of Urban Economics*, 144, 103714.

Hierons, T. (2025): "Optimal Congestion Pricing in General Equilibrium," Doctoral dissertation.

Hintermann, B., B. Schoeman, J. Molloy, T. Götschi, A. Castro, C. Tchervenkov, U. Tomic, and K. W. Axhausen (2026): "Pigovian Transport Pricing in Practice," *The Review of Economic Studies*.

Imbens, G. W., P. Johnson, and R. H. Spady (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333–357.

Ireland, C. T. and S. Kullback (1968): "Contingency tables with given marginals," *Biometrika*, 55, 179–188.

Jaynes, E. T. (1957): "Information theory and statistical mechanics," *Physical review*, 106, 620.

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.

Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica: Journal of the Econometric Society*, 861–874.

Knight, F. H. (1924): "Some Fallacies in the Interpretation of Social Cost," *The Quarterly Journal of Economics*, 38, 582–606.

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

Kreindler, G., A. Gaduh, T. Graff, R. Hanna, and B. A. Olken (2023): "Optimal Public Transportation Networks: Evidence from the World's Largest Bus Rapid Transit System in Jakarta," NBER Working Paper 31369, National Bureau of Economic Research.

Kullback, S. (1997): *Information theory and statistics*, Courier Corporation.

Kullback, S. and R. A. Leibler (1951): "On information and sufficiency," *The annals of mathematical statistics*, 22, 79–86.

Leape, J. (2006): "The London Congestion Charge," *Journal of Economic Perspectives*, 20, 157–176.

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.

Lieberman, G. (2021): "Israel's traffic jams are back, and worse than ever," *Globes*.

Martin, L. A. and S. Thornton (2017): "Can Road Charges Alleviate Congestion?" *Available at SSRN 3055522*.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619.

National Insurance Institute of Israel (2026): "Minimum Wage — General Information," .

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.

Severen, C. (2023): "Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification," *The Review of Economics and Statistics*, 105, 1073–1091.

——— (2026): "Evaluating Transportation Improvements Within Cities Using Quantitative Spatial Models," Working Paper 26-13, Federal Reserve Bank of Philadelphia.

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

Small, K., E. Verhoef, and R. Lindsey (2024): *The Economics of Urban Transportation*, Taylor & Francis.

Small, K. A. (1982): "The Scheduling of Consumer Activities: Work Trips," *The American Economic Review*, 72, 467–479.

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

Small, K. A., C. Winston, and J. Yan (2005): "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability," *Econometrica*, 73, 1367–1382.

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.

Tarduno, M. (2024): "For Whom the Bridge Tolls: Congestion, Air Pollution, and Second-Best Road Pricing," Energy Institute Working Paper WP-323R, Energy Institute at Haas, University of California, Berkeley.

The Times of Israel (2020): "Torrential rains in northern Israel break 51-year record," *The Times of Israel*.

——— (2021): "Winter carved country into two regions: Above-average rain, and drought," *The Times of Israel*.

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

Tsivanidis, N. (2026): "Evaluating the Impact of Urban Transit Infrastructure: Evidence from Bogotá's TransMilenio," *American Economic Review*, 116, 418–463.

U.S. Census Bureau (2022): "American Community Survey, 2022 1-Year Estimates," Tech. rep., U.S. Census Bureau.

U.S. Department of Transportation, Federal Highway Administration (2017): "National Household Travel Survey," Tech. rep., Federal Highway Administration, 2022 figures based on the Next Generation National Household Travel Survey.

van den Berg, V. A. C. and E. T. Verhoef (2011): "Winning or Losing from Dynamic Bottleneck Congestion Pricing? The Distributional Effects of Road Pricing with Heterogeneity in Values of Time and Schedule Delay," *Journal of Public Economics*, 95, 983–992.

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

Walters, A. A. (1961): "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, 29, 676–699.

Wikipedia (2026): "Tel Aviv Metro," https://en.wikipedia.org/wiki/Tel_Aviv_Metro, accessed: 2026-07-15.

Yang, J., A.-O. Purevjav, and S. Li (2020): "The marginal cost of traffic congestion and road pricing: evidence from a natural experiment in Beijing," *American Economic Journal: Economic Policy*, 12, 418–453.

Zagrizak, A. (2025): "Average peak-hour speed on Ayalon Highway just 14 km/h," *Globes*.

A Supplemental Tables and Figures

A.1 Descriptive Statistics

			Quantiles				
Variable	Mean	SD	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

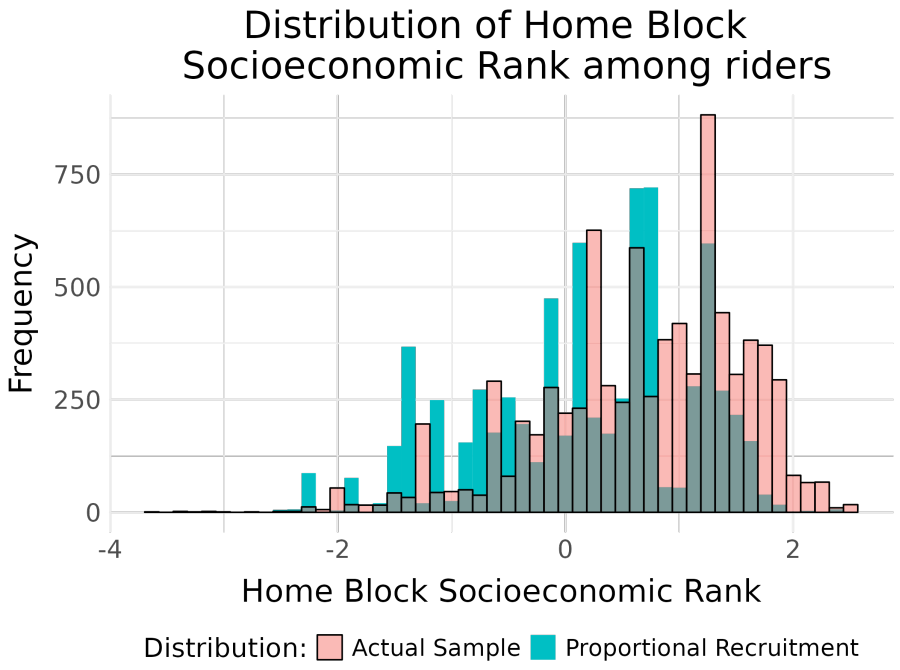
Appendix Table A1: Summary Statistics of Main Outcome Variables, Pre-Treatment

			Quantiles				
Covariate	Mean	SD	10	25	50	75	90
Congestion Covariates							
Price / Trip	3.41	3.14	0.29	1.06	2.63	4.75	7.62
Price / Km	0.30	0.25	0.04	0.10	0.23	0.43	0.65
# Cross-City Trips	5.99	4.04	2.00	2.50	5.00	9.00	11.00
# Within-City Trips	4.02	3.92	1.00	2.00	3.00	5.00	8.50
Flexibility Covariates							
Last AM Arrival Time	9.48	1.23	7.73	8.67	9.76	10.46	10.80
Last AM Arrival Entropy	1.41	0.45	0.69	1.10	1.39	1.73	1.95
First PM Departure Time	17.08	0.90	16.11	16.42	16.89	17.54	18.29
First PM Departure Entropy	1.81	0.46	1.10	1.59	1.89	2.15	2.35
Socioeconomic Covariates							
Home Block Socioeconomic Rank	0.68	0.90	-0.62	0.16	0.82	1.33	1.75
Home Block Population Density	0.01	0.01	0.00	0.00	0.01	0.01	0.02
Route-Based Covariates							
Med # Trips / Week with Metro	7.96	7.49	0.00	1.50	6.50	13.00	18.00
Med # Transitions b/w Zones	2.81	3.73	0.00	0.00	1.50	4.00	7.50
Med Metro Distance in AM Peak	12.85	20.23	0.00	0.00	4.00	17.71	38.87
Med Metro Distance in PM Peak	19.08	19.62	0.00	5.13	14.34	26.60	43.30

Appendix Table A2: Summary Statistics of Heterogeneity Covariates, Pre-Treatment. Entropy ($-\sum_b p_b \log p_b$) is computed after rounding times to 10-minute bins. Higher values indicate less regular schedules; lower values indicate more routine timing. Clock-time covariates (last AM arrival, first PM departure) are in hours of the day. N = 7710 is the number of drivers with cross-city trips in the heterogeneity analysis sample (Figure 7). Quantiles are computed across riders.

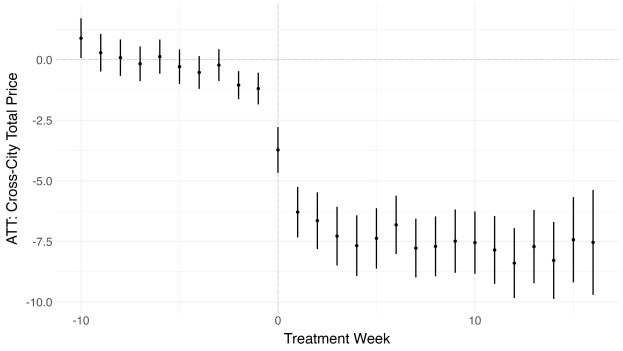
			Quantiles				
Covariate	Mean	SD	10	25	50	75	90
Last AM Arrival Time	9.48	1.24	7.73	8.67	9.76	10.46	10.80
Last AM Arrival Entropy	1.19	0.51	0.64	0.87	1.22	1.56	1.79
First PM Departure Time	17.08	0.91	16.11	16.43	16.89	17.55	18.30
First PM Departure Entropy	1.60	0.53	0.87	1.33	1.67	1.98	2.23
Public Transit 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

Appendix Table A3: Summary statistics of Pre-treatment Driver Characteristics

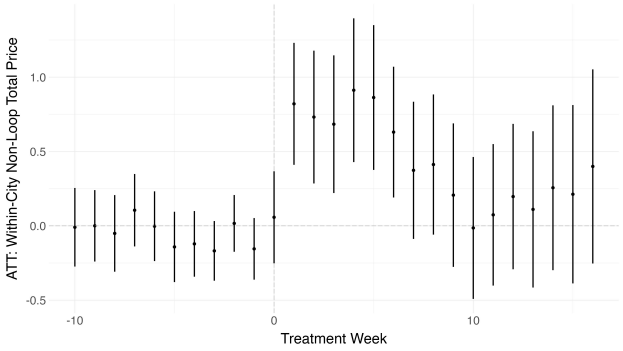


Appendix Figure A1: Socioeconomic Index Distribution in Sample and in Population

A.2 Additional Event Studies

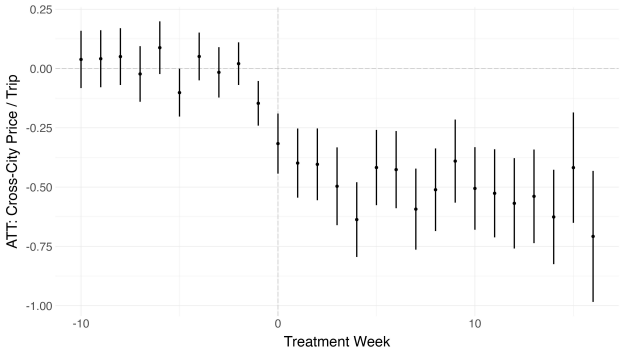


(a) Outcome: total price of cross-city trips per week

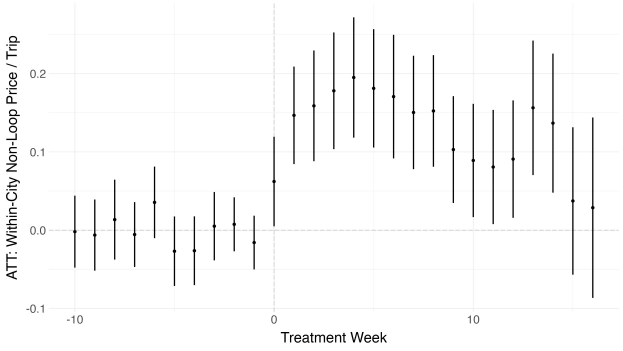


(b) Outcome: total price of within-city non-loop trips per week

Appendix Figure A2: Event studies of dynamic average treatment effects across weeks relative to treatment week.

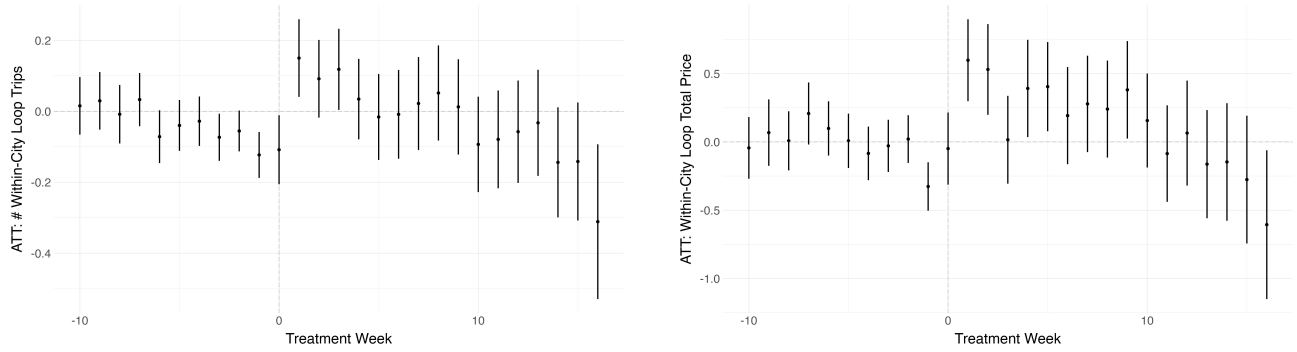


(a) Outcome: price per trip of cross-city trips per week



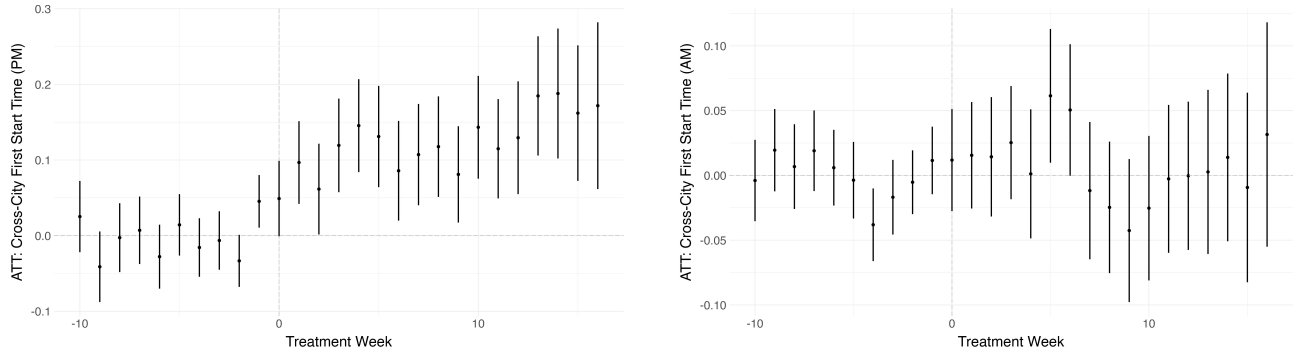
(b) Outcome: price per trip of within-city non-loop trips per week

Appendix Figure A3: Event studies of dynamic average treatment effects across weeks relative to treatment week.



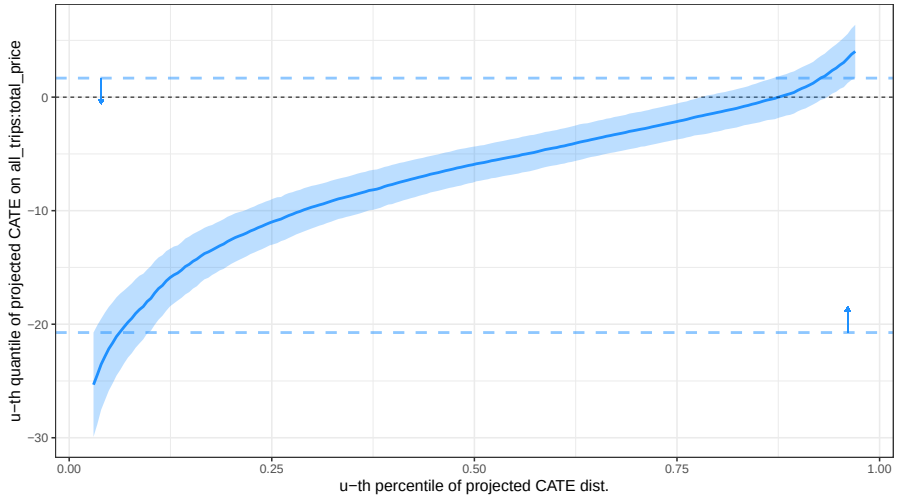
(a) Outcome: number of within-city loop trips per week *(b) Outcome: total price of within-city loop trips per week*

Appendix Figure A4: Event studies of dynamic average treatment effects across weeks relative to treatment week.



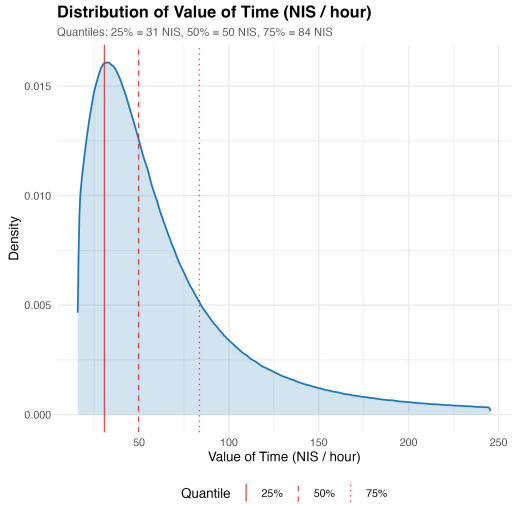
(a) Outcome: first PM trip start time for cross-city trips *(b) Outcome: first AM trip start time for cross-city trips*

Appendix Figure A5: Event studies of dynamic average treatment effects on departure times for cross-city trips.



Appendix Figure A6: 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.

A.3 Demand Estimation



Appendix Figure A7: The estimated distribution of drivers' Value of Time, γ_i/α_i.

Best modes (median start time)	N	Share	P10	Median	P90	Mean
1x BUS	24,319	42.8%	0.39	0.71	1.33	0.83
2x BUS	13,246	23.3%	0.38	0.66	1.10	0.72
Other (mixed walk+bus)	7,953	14.0%	0.40	0.78	1.67	0.98
WALK ONLY	4,954	8.7%	0.28	0.57	1.38	0.77
3x BUS	4,715	8.3%	0.38	0.64	1.01	0.67
4+ x BUS	1,136	2.0%	0.39	0.62	0.93	0.64
Has RAIL	543	1.0%	0.36	0.62	0.96	0.66

Appendix Table A4: Distribution of imputed driving / best transit duration by best-modes bucket. Snapshot at each trip's median start time, one row per trip id.

A.4 Counterfactual Equilibria

	Moments		*Percentiles*				
	Mean	SD	P20	P40	P50	P60	P80
All times							
Length-weighted	86.5	21.1	74.0	79.9	86.3	89.6	105.4
Occupancy-weighted	35.0	38.6	1.7	5.2	8.6	31.1	78.9
Peak							
Length-weighted	82.2	25.3	69.8	78.0	80.9	87.3	100.9
Occupancy-weighted	29.0	36.3	1.5	4.2	5.9	10.9	74.0
Off-peak							
Length-weighted	87.6	19.5	76.4	79.9	86.9	89.6	105.2
Occupancy-weighted	43.1	39.8	2.1	9.6	27.0	69.2	83.7
Leisure (unpriced)							
Length-weighted	90.5	16.2	79.7	80.0	89.3	90.0	109.0
Occupancy-weighted	60.4	39.8	4.6	68.8	78.1	84.3	93.1

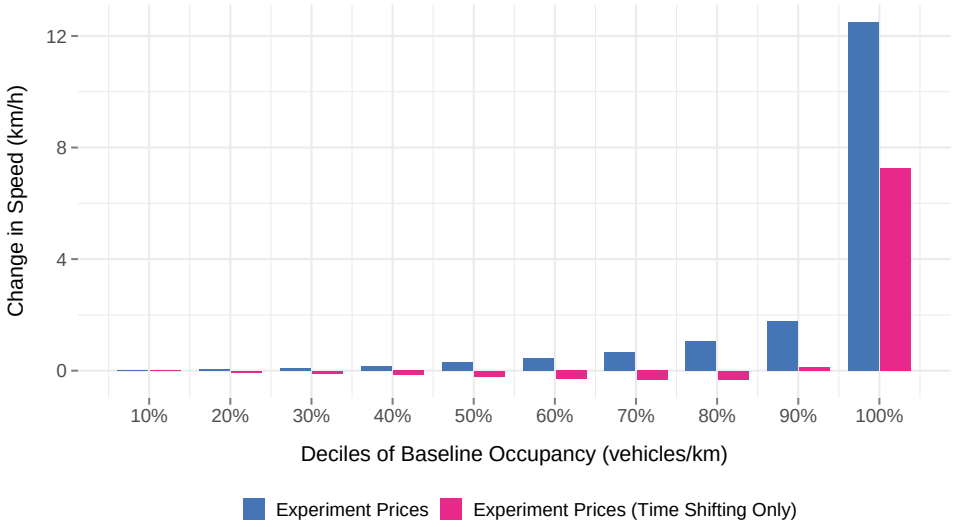
Appendix Table A5: Baseline (no-price) equilibrium segment speeds (km/h): mean, SD, and quintiles; highways only (n = 32,576; peak n = 10,180; off-peak n = 15,270; leisure n = 7,126)

	Moments		*Percentiles*				
	Mean	SD	P20	P40	P50	P60	P80
Consumer Surplus (NIS/day)	2.80	77.85	1.95	5.48	7.25	9.39	16.73
Trip Probability (%)	32.3	25.9	9.9	18.5	24.1	31.1	53.8
Trip Duration (mins)	21.4	19.3	7.6	12.9	15.5	19.6	31.5
Trip Distance (km)	15.40	18.91	5.18	8.09	10.03	12.58	19.63
Trip Speed (km/h)	43.6	15.9	28.3	40.4	45.1	48.1	56.4

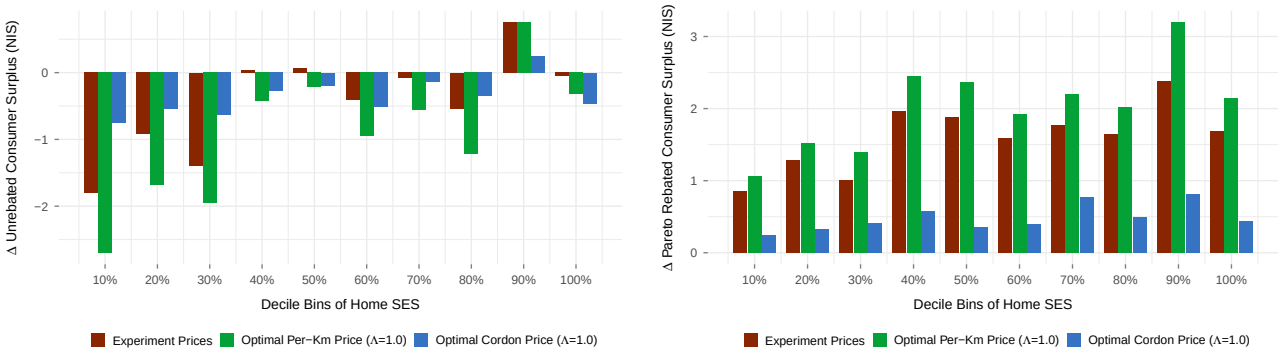
Appendix Table A6: Baseline (no-price) equilibrium trip-level outcomes: population-weighted mean, SD, and quintiles across consumer types (types = 733,802; population mass = 2,770,224)

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	0.00	0.00	-0.00	0.00	0.00
Distance / Day (km)	0.00	0.00	-0.00	0.00	0.00
Price / Trip (NIS)	6.56	7.65	1.30	4.11	9.06
Duration / Trip (mins)	-2.69	5.22	-3.40	-0.27	0.01
Speed / Trip (km/h)	3.75	6.21	-0.04	0.91	6.52
Raw Consumer Surplus (NIS/day)	-1.71	6.55	-2.16	-0.47	-0.04
Rebated Consumer Surplus (NIS/day)	0.87	6.55	0.41	2.11	2.53

Appendix Table A7: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the experiment fee schedule with time-shifting choice only (trip probabilities held fixed), relative to the baseline equilibrium with no pricing.



Appendix Figure A8: Average change in equilibrium segment-15-minute interval speeds due to the experiment fee schedule, by baseline occupancy deciles, with and without the extensive margin (time shifting only).



(a) Unrebated *(b) Pareto-rebated*

Appendix Figure A9: Distribution of per-driver welfare changes by socioeconomic-status decile under different pricing schemes.

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	-8.88	15.94	-13.54	-4.89	-0.45
Distance / Day (km)	-1.91	6.63	-1.74	-0.41	-0.02
Price / Trip (NIS)	11.75	11.30	4.41	8.68	16.31
Duration / Trip (mins)	-4.77	6.97	-6.88	-1.68	-0.08
Speed / Trip (km/h)	8.16	9.10	0.44	4.91	13.21
Raw Consumer Surplus (NIS/day)	-0.93	10.35	-2.24	-0.68	-0.06
Rebated Consumer Surplus (NIS/day)	2.02	10.35	0.72	2.27	2.89
Pareto-Rebated Consumer Surplus (NIS/day)	2.02	9.26	0.82	0.82	0.82

Appendix Table A8: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the welfare-optimal usage-based (per-km) price schedule ($\lambda = 1$), relative to the baseline equilibrium with no pricing.

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	-2.57	9.63	-0.10	0.02	0.55
Distance / Day (km)	-0.52	2.77	-0.01	0.00	0.07
Price / Trip (NIS)	3.79	8.66	0.00	0.00	0.00
Duration / Trip (mins)	-1.10	2.42	-1.00	-0.15	-0.00
Speed / Trip (km/h)	1.77	4.01	0.03	0.37	1.46
Raw Consumer Surplus (NIS/day)	-0.37	3.73	-0.02	0.00	0.12
Rebated Consumer Surplus (NIS/day)	0.48	3.73	0.82	0.85	0.96
Pareto-Rebated Consumer Surplus (NIS/day)	0.48	2.52	0.11	0.11	0.22

Appendix Table A9: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the welfare-optimal cordon price schedule ($\lambda = 1$), relative to the baseline equilibrium with no pricing.

Pricing scheme	Avg. revenue (NIS/day per capita)	Avg. compensation (NIS/day per capita)	Avg. remainder rebate (NIS/day per capita)	Share of drivers compensated (%)
Optimal usage-based (λ=1)	2.955	2.139	0.816	80.4%
Optimal cordon (λ=1)	0.843	0.738	0.106	34.9%
Experiment prices (λ=1)	2.044	1.401	0.644	74.2%

Appendix Table A10: Feasibility of a Pareto-improving lump-sum rebate under λ=1 pricing.

Notes. All monetary figures are mass-weighted averages in NIS/day per capita. "Avg. compensation" is the per-capita revenue needed to hold every losing driver type at its baseline surplus (the mean of per-type losses $\max\{0, CS^{\text{base}} - CS^{\text{cf}}\}$). "Avg. remainder rebate" is the equal per-capita rebate of the leftover revenue (average revenue minus average compensation); it is what every driver receives on top of any compensation. "Share of drivers compensated" is the population mass with a positive CS loss. Because average revenue exceeds average compensation in all three schemes, the remainder rebate is positive, so after compensating losers every driver type is weakly better off than the baseline.

	Peak	Off-peak
Metropolis	0.00	0.00
Metro Fringe	0.00	0.00
Periphery	2.34	0.00

Appendix Table A11: Optimal usage-based price levels with no revenue recycling ($\Lambda = 0$).

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	0.02	2.62	-0.01	0.00	0.05
Distance / Day (km)	-0.10	2.39	-0.00	0.00	0.00
Price / Trip (NIS)	0.20	2.95	0.00	0.00	0.00
Duration / Trip (mins)	-0.15	0.76	-0.04	-0.00	0.00
Speed / Trip (km/h)	0.18	0.98	-0.01	0.00	0.08
Raw Consumer Surplus (NIS/day)	0.03	1.03	-0.00	0.00	0.01
Rebated Consumer Surplus (NIS/day)	0.06	1.03	0.03	0.03	0.04
Pareto-Rebated Consumer Surplus (NIS/day)	0.06	0.63	-0.04	-0.04	-0.03

Appendix Table A12: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the optimal usage-based (per-km) price schedule ($\lambda = 0$), relative to the baseline equilibrium with no pricing.

Zone / crossing × time	Estimated	Homogeneous
Usage-based schedule		
Metro Fringe, peak	0.23	0.56
Metro Fringe, off-peak	0.37	0.84
Metropolis, peak	0.11	**0.44**
Metropolis, off-peak	0.38	0.71
Periphery, peak	**0.00**	0.70
Periphery, off-peak	0.04	1.13
Cordon schedule		
Metro Fringe → Metropolis, peak	0.26	**0.58**
Metro Fringe → Metropolis, off-peak	0.48	0.77
Metropolis → Metro Fringe, peak	0.23	0.59
Metropolis → Metro Fringe, off-peak	0.37	0.75
Metro Fringe → Periphery, peak	**0.00**	0.64
Metro Fringe → Periphery, off-peak	0.33	1.11
Periphery → Metro Fringe, peak	0.61	0.76
Periphery → Metro Fringe, off-peak	0.14	0.61

Appendix Table A13: Per-bin first-order revenue recycling thresholds under usage-based and cordon pricing. Bold entries mark the binding bin whose threshold equals the overall Λ^* for that scheme and demand specification.

(a) Usage-based

	Peak	Off-peak
Metropolis	1.15	0.51
Metro Fringe	0.22	0.05
Periphery	0.02	0.00

(b) Cordon-based

	Peak	Off-peak
Metropolis → Metro Fringe	5.25	3.64
Metro Fringe → Metropolis	6.42	5.52
Metro Fringe → Periphery	0.48	0.00
Periphery → Metro Fringe	0.09	0.09

Appendix Table A14: Optimal price levels under homogeneous demand with full revenue recycling ($\Lambda = 1$), for usage-based and cordon-based pricing. Demand parameters are fixed at the population average for all drivers.

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	-5.71	8.07	-8.40	-3.27	-0.56
Distance / Day (km)	-1.08	2.70	-1.03	-0.27	-0.04
Price / Trip (NIS)	7.47	6.46	2.83	5.68	10.36
Duration / Trip (mins)	-3.25	4.82	-4.86	-1.03	-0.05
Speed / Trip (km/h)	5.33	5.93	0.25	3.14	8.83
Raw Consumer Surplus (NIS/day)	-1.48	2.92	-1.61	-0.47	-0.06
Rebated Consumer Surplus (NIS/day)	0.71	2.92	0.57	1.71	2.12
Pareto-Rebated Consumer Surplus (NIS/day)	0.71	0.27	0.65	0.65	0.65

Appendix Table A15: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the usage-based (per-km) price schedule optimized for homogeneous demand ($\lambda = 1$), relative to the baseline equilibrium with no pricing computed under the same homogeneous-demand assumptions.

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	-1.51	5.24	-0.03	0.00	0.17
Distance / Day (km)	-0.26	1.14	-0.00	0.00	0.02
Price / Trip (NIS)	1.92	4.26	0.00	0.00	0.00
Duration / Trip (mins)	-0.49	1.23	-0.37	-0.05	-0.00
Speed / Trip (km/h)	0.80	2.01	0.01	0.15	0.69
Raw Consumer Surplus (NIS/day)	-0.36	1.54	-0.00	0.00	0.03
Rebated Consumer Surplus (NIS/day)	0.10	1.54	0.45	0.46	0.49
Pareto-Rebated Consumer Surplus (NIS/day)	0.10	0.18	0.04	0.04	0.07

Appendix Table A16: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the cordon price schedule optimized for homogeneous demand ($\lambda = 1$), relative to the baseline equilibrium with no pricing computed under the same homogeneous-demand assumptions.

	Mean	SD	Q25	Median	Q75
Trip Probability (%)	-5.28	5.60	-7.93	-3.86	-1.50
Distance / Day (km)	-0.81	0.86	-1.22	-0.59	-0.23
Price / Trip (NIS)	7.10	0.00	7.10	7.10	7.10
Duration / Trip (mins)	-2.35	0.11	-2.43	-2.34	-2.27
Speed / Trip (km/h)	4.99	0.16	4.88	4.99	5.11
Raw Consumer Surplus (NIS/day)	-0.57	1.31	-0.96	-0.48	-0.21
Rebated Consumer Surplus (NIS/day)	0.70	1.31	0.31	0.79	1.06
Pareto-Rebated Consumer Surplus (NIS/day)	NaN	NaN	NA	NA	NA

Appendix Table A17: Changes in driver-level trip outcomes and money-metric welfare in the equilibrium under the welfare-optimal price ($\lambda = 1$) in the counterfactual with no network heterogeneity, relative to the baseline equilibrium with no pricing computed under the same no-network-heterogeneity assumptions.

B Motivating Example Derivations

This appendix derives the local welfare decomposition in Section 2 directly from the welfare objective. Recall that

$$\delta_i(p) \coloneqq V_i - \gamma_i D(q^*(p)) - \alpha_i p, \qquad q_i^*(p) \coloneqq F(\delta_i(p)), \qquad q^*(p) \coloneqq \sum_i q_i^*(p),$$

and the welfare objective is

$$W_\Lambda(p) \coloneqq CS(p) + \Lambda R(p) = \sum_i \frac{1}{\alpha_i} \left[q_i^*(p)\delta_i(p) + \mathcal{K}(q_i^*(p)) \right] + \Lambda p q^*(p). \tag{12}$$

Since $q_i^*(p) = F(\delta_i(p))$, the equilibrium choice probability solves the following individual problem:

$$q_i^*(p) \in \arg\max_{x \in [0,1]} \left\{ x\delta_i(p) + \mathcal{K}(x) \right\}. \tag{13}$$

The first-order condition for Equation (13) is therefore

$$\delta_i(p) + \mathcal{K}'(q_i^*(p)) = 0. \tag{14}$$

Differentiating Equation (12) along the equilibrium path gives

$$\frac{\mathrm{d}W_\Lambda}{\mathrm{d}p} = \sum_i \frac{1}{\alpha_i} \left[\left(\delta_i(p) + \mathcal{K}'(q_i^*(p)) \right) \frac{\mathrm{d}q_i^*(p)}{\mathrm{d}p} + q_i^*(p) \frac{\mathrm{d}\delta_i(p)}{\mathrm{d}p} \right] + \Lambda \left(q^*(p) + p \frac{\mathrm{d}q^*(p)}{\mathrm{d}p} \right).$$

By Equation (14), the first term inside brackets is zero. Moreover,

$$\frac{\mathrm{d}\delta_i(p)}{\mathrm{d}p} = -\gamma_i \frac{\mathrm{d}D(q^*(p))}{\mathrm{d}p} - \alpha_i.$$

Hence, defining the value of time as $\phi_i \coloneqq \gamma_i / \alpha_i$, we have

$$\frac{\mathrm{d}W_\Lambda}{\mathrm{d}p} = -\frac{\mathrm{d}D(q^*(p))}{\mathrm{d}p} \sum_i \phi_i q_i^*(p) - (1-\Lambda) q^*(p) + \Lambda p \frac{\mathrm{d}q^*(p)}{\mathrm{d}p}. \tag{15}$$

At $p = 0$, the final term in Equation (15) drops out. Writing

$$\dot{D}(0) \coloneqq \frac{\mathrm{d}}{\mathrm{d}p} D(q^*(p)) \bigg|_{p=0},$$

we obtain

$$\frac{\mathrm{d}W_\Lambda}{\mathrm{d}p} \bigg|_{p=0} = -\dot{D}(0) \sum_i \phi_i q_i^*(0) - (1-\Lambda) q^*(0).$$

Finally, since $q^*(0) = \sum_i q_i^*(0)$ and $\bar{\phi} \coloneqq \sum_i \phi_i$,

$$\sum_i \phi_i q_i^*(0) = q^*(0)\bar{\phi} + \mathrm{Cov}(q_i^*(0),\, \phi_i)\,. \tag{16}$$

Substituting Equation (16) gives

$$\left.\frac{\mathrm{d}W_\Lambda}{\mathrm{d}p}\right|_{p=0} = \underbrace{-\dot{D}(0)q^*(0)\bar{\phi}}_{\text{average congestion relief}} \underbrace{-\dot{D}(0)\,\mathrm{Cov}(q_i^*(0),\, \phi_i)}_{\text{sorting gain}} - \underbrace{(1-\Lambda)q^*(0)}_{\text{revenue leakage}}.$$

C Causal Effect Estimation Details

C.1 Identification and Estimation of Average Treatment Effects

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. (2024) 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 $Y_{it} \coloneqq Y_{it}(C_i \mathbb{1}\{C_i < t\})$ 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 $c \leq t$:

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

To identify CATTs τ_{ct}, we impose a two-way fixed effects (TWFE) model of control potential outcomes as in Borusyak et al. (2024):

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

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., 2024).

To translate the difference-in-differences identification intuition into an estimation strategy that will also allow us to study treatment effect heterogeneity in Section 4.2, we follow the imputation approach of Borusyak et al. (2024). 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:

$$(\hat{\alpha}_i, \hat{\gamma}_t) := \underset{\alpha_i, \gamma_t}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{C_i > t, C_i - M_i \geq t\} \left(Y_{it} - [\alpha_i + \gamma_t]\right)^2. \tag{17}$$

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., 2024). 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:[31]

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

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:[32] $\hat{\tau} := \frac{1}{12} \sum_{e=4}^{16} \hat{\tau}_e$.

[31] 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.

[32] As we show in Section 4.1, our event study results suggest that participants adjusted their behavioral responses in

We discuss inference for $\hat{\tau}$ in Section 4.1.

C.2 Characterizing Treatment Effect Heterogeneity via Sorted Effects

The control potential outcome model used to identify treatment effects in Section 4.1 allows for constructing noisy estimates of 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 Appendix Section C.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 4.1. 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:[33]

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

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 in Equation (17) 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 Appendix Figure A6 in Appendix Section A.2. Since the uniform confidence band does not contain any horizontal lines,[34] 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.

We next 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

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.

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

[34] We can see that no horizontal lines pass through the uniform confidence band since the dashed line in Appendix Figure A6 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.

treatment effects:

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

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

D Habitual Trip Panel Construction

This appendix describes how we construct the panel used to estimate the demand model in Section 6.1. For each driver, the panel contains potential and realized trips over time, together with their durations, prices, and public transit alternatives across potential departure times each day.

D.1 Defining Habitual Trips

As introduced in Section 3.3 and discussed in Section 6.1, we estimate demand for recurring habitual trips, such as commutes to and from work, rather than one-off trips such as infrequent medical visits.

We begin with 23.4 million travel records and retain trips that occur in the afternoon or evening (2:30 PM to 11 PM), on weekdays (Sunday through Thursday in Israel), and on days without a COVID-19 lockdown. We also require the origin and destination to lie in different TAZs. We then define effective origin–destination pairs: for trips shorter than 9.66 kilometers, an origin–destination pair consists of the exact pair of trip endpoint TAZs present in the data. For longer trips, we cluster neighboring TAZs at each endpoint to accommodate small spatial differences, such as variation in parking locations.

We then identify habitual trips in the filtered data using the following algorithm:

1. Bin trips by driver and origin–destination pair.
2. Move an eight-week rolling window through the sample one week at a time. For each bin observed in the window:
   (a) Define reference medians for *distance* and *start time* within the window. If the bin formed a habitual trip in an earlier window, retain the existing medians.

(b) Classify a trip as *close* when its distance differs from the reference median by less than 30% and its start time differs from the reference median by less than one hour.

(c) If more than two of the eight weeks in the current window contain at least one close trip, assign the habitual-trip label to qualifying trips or extend an existing label.

3. After assigning labels, discretize typical start times into 15-minute bins.

We restrict the resulting sample to habitual trips observed through the end of June 2021 whose median start time falls between 3 PM and 8 PM on non-lockdown days. The demand-estimation sample contains 68,562 habitual trips, corresponding to 521,646 realized trips. It covers 7,350 of the 10,119 individuals with at least one evening habitual trip.

Given these habitual trips, we construct our panel of opportunities to depart on these consistent trips as follows. For each week in which a habitual trip is active and each weekday (Sunday through Thursday), we define nine potential departure times: the habitual trip's median departure time and four earlier and four later times in 15-minute increments. We label options observed in the raw data as "realized"; the remaining options represent departure times the driver could have chosen. The demand model uses this variation to estimate preferences over departure times, trip durations, and trip prices.

We next impute a route for each habitual trip. As described in Section 3.1, in the raw data, we only observe a subset of the TAZs that a driver traverses during their trip.[35] Therefore, we reconstruct a complete sequence of road segments in the OpenStreetMap (OSM) network introduced in Section 3.2 that are consistent with the observed TAZ sequence. These routes allow us to estimate the duration and price a driver would face at different departure times in Appendix Section D.3 and to identify the road segments used in the counterfactuals; see Appendix Section F.2.

The route-imputation procedure uses dynamic programming to find the globally fastest feasible path consistent with the observed sequence of TAZs. Let $(O, I_1, I_2, \ldots, I_N, D)$ denote a route-finding problem, where O and D are the origin and destination and $I_1, \ldots, I_N$ are the observed intermediate TAZs. A feasible route must traverse the intermediate TAZs in the observed order.

First, the algorithm precomputes the fastest paths between entry nodes of TAZs that can occupy adjacent positions in the sequence. An *entry node* is an OSM node on the boundary of a TAZ. The algorithm stores the resulting paths and travel times for reuse across route-finding problems. To accommodate repeated TAZs in the observed sequence, the algorithm partitions the sequence into consecutive subsequences containing no duplicate TAZs. Within each subsequence, it constructs a graph whose edges represent the precomputed fastest paths between entry nodes of adjacent constraints, with edge weights equal to travel times. It then computes the shortest path from the origin, or from each feasible entry node of the preceding subsequence, to every entry node of the final TAZ in the current subsequence. These shortest-path results are carried forward to the next subsequence. The algorithm repeats this procedure until it reaches the destination. The resulting route minimizes total travel time among paths that traverse the observed intermediate TAZs in the required order.

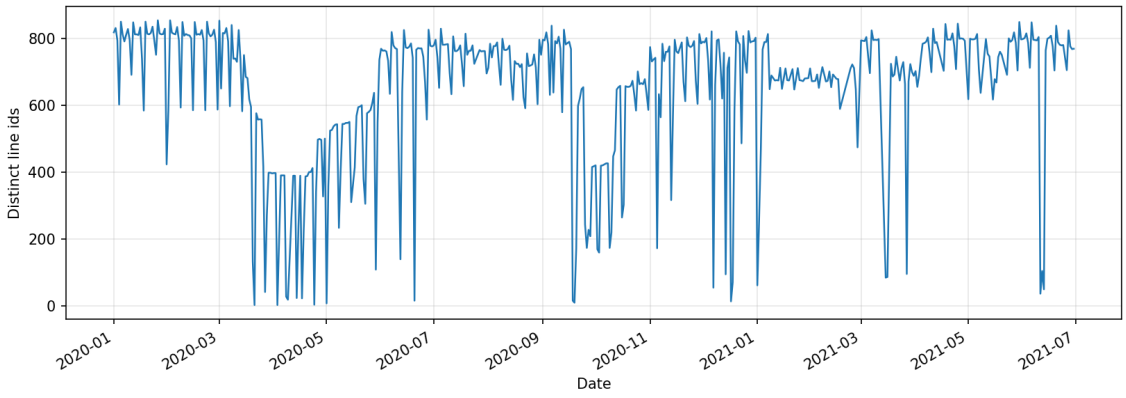
[35] Recall that, for privacy reasons, we observe a TAZ visited during a trip only when the driver crosses a spatial or temporal pricing-zone boundary.

D.2 Constructing Segment-Level Bus Speeds

To predict potential trip durations at different departure times, we construct bus-speed estimates for every OpenStreetMap (OSM) road segment, 15-minute time bin, and day in our habitual trip panel. The resulting road speed panel covers approximately 2.3 billion road-kilometer–15-minute cells, including 0.6 billion cells corresponding to intervals between 3 PM and 9 PM. We construct the road speed panel in three stages: first, we match official bus routes to the OSM network; then we estimate road speeds directly from bus GPS pings; and finally, we impute speeds for segment–time cells without direct bus coverage.

The construction combines three inputs. First, we observe 1.4 billion GPS pings from 925 bus lines, including each bus's location, timestamp, and trip identifier. Second, we use official transit-route shape files from the Israeli National Authority for Public Transport (NAPT). Third, we use the 2021 OSM road network for Israel. Appendix Figure D1 shows the number of active bus lines observed each day.

Appendix Figure D1: Active bus lines observed by date



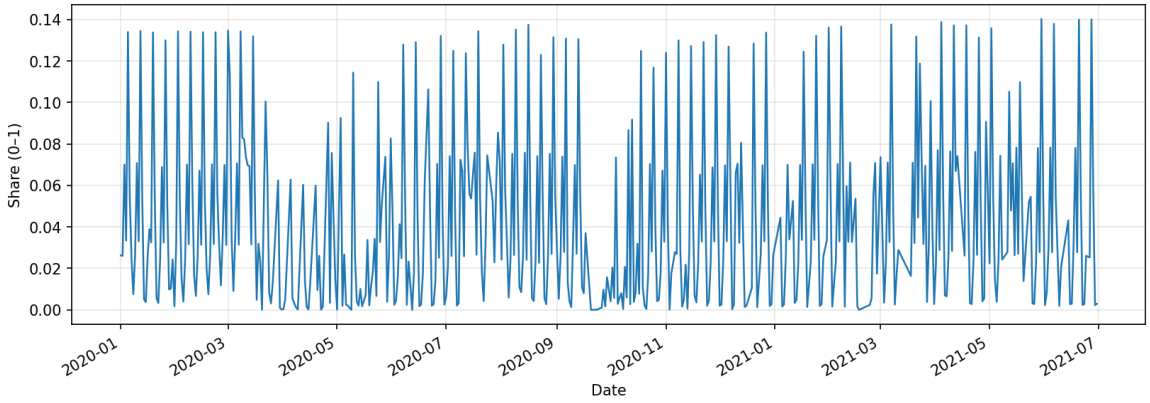
Route Matching Because the official route shapes are not defined on the OSM network, we first map each bus route to an ordered path on the OSM road graph. Geometric intersections alone are unreliable: they can include nearby roads incorrectly, handle loops poorly, and confuse the direction of travel. We therefore shorten candidate OSM segments at both endpoints to reduce spurious matches with nearby intersecting roads. We then retain segments whose direction agrees with the direction implied by the route shape and connect the matched segments using shortest-path queries on the local OSM graph. This procedure produces an ordered, directed OSM route for each bus line.

Direct Speed Estimation We next use consecutive GPS pings to infer bus movement along the OSM segments between them. For each pair of pings, we calculate predicted free-flow travel time from the speed limits of the OSM segments traversed along the bus route between those pings and define a "travel-time factor" as realized travel time divided by predicted free-flow travel time. This common factor accommodates differences in speed limits across the traversed OSM segments. For each traversed OSM edge, we obtain the implied bus speed by dividing the edge's speed limit by the

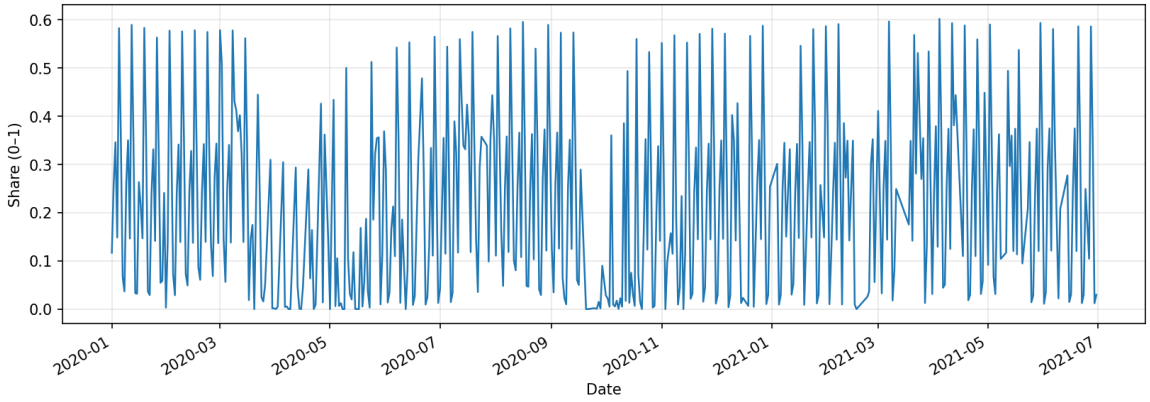
travel-time factor. If a segment has multiple observations within a 15-minute bin, we average them using the share of the segment traversed as weights.

This procedure yields direct speed estimates for many segment–time cells, but coverage remains incomplete. Appendix Figure D2 shows the daily share of road-network kilometer-hours covered by direct bus-speed estimates. Coverage ranges from 7% to 14% across the full Israeli road network. Coverage is substantially higher on major highways during the period of interest as visualized in Appendix Figure D4, often reaching approximately 60% as shown in Appendix Figure D3.

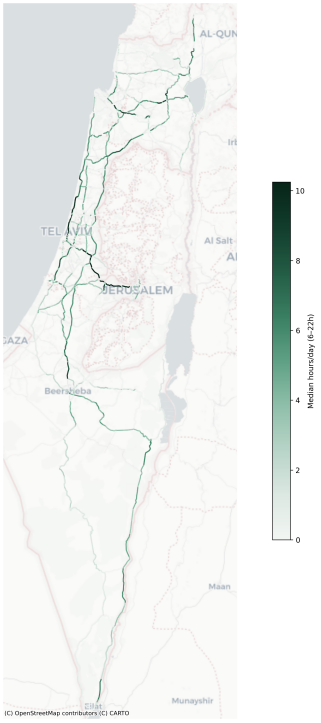
Appendix Figure D2: Road-network kilometer-hour coverage: all roads, full day



Appendix Figure D3: Road-network kilometer-hour coverage: major roads, 6 AM–10 PM



Appendix Figure D4: Median hours of daily coverage: major roads, 6 AM–10 PM



Imputing Missing Speeds Because the direct road speed estimates cover only a subset of segment–time cells, we impute the remaining speeds using a sequence of fixed-effects regressions estimated separately for each calendar day. In directly observed cells, bus speed equals the segment's speed limit divided by the estimated travel-time factor. The remaining cells initially contain missing values. Each segment r has two time-invariant attributes from the OSM road network that we use for speed imputation: $z(r)$ denotes the segment's *pricing zone*: Metropolis, Metropolitan fringe, or Periphery; and $\ell(r)$ denotes the segment's OSM *road type*, given by its `highway` tag, such as motorway, trunk, primary, secondary, or residential.

Let y_{rt} denote bus speed in kilometers per hour on road segment r during 15-minute time bin t. We estimate a sequence of fixed-effects regressions of the form

$$y_{rt} = \mu_{rt} + \varepsilon_{rt},$$

where μ_{rt} denotes the sum of a set of fixed effects for a given specification. At each step of the imputation pipeline, we estimate the regression using all currently nonmissing observations; we then sum the estimated fixed effects to predict y_{rt} for missing cells whose fixed-effect levels were observed in estimation. The next specification is applied to cells that remain missing.

We apply the following eleven specifications for μ_{rt} in order:

1. $\mu_{rt} = \xi_r + \alpha_{z(r)\ell(r)t}$: a segment fixed effect and a zone-by-road-type-by-time-bin interaction. This specification combines each segment's permanent speed level with the time-varying congestion pattern for its zone-by-road-type group;

2. $\mu_{rt} = \xi_r + \alpha_{\ell(r)t}$: a segment fixed effect and a road-type-by-time-bin interaction, omitting zone from the time-varying component;

3. $\mu_{rt} = \xi_r + \alpha_{z(r)t}$: a segment fixed effect and a zone-by-time-bin interaction, omitting road type from the time-varying component;

4. $\mu_{rt} = \xi_r + \alpha_t$: a segment fixed effect and a time-bin fixed effect. The time-varying component is common across zones and road types;

5. $\mu_{rt} = \xi_r$: a segment fixed effect only, which assigns each remaining cell the segment's average daily speed without time variation;

6. $\mu_{rt} = \alpha_{z(r)\ell(r)t}$: a zone-by-road-type-by-time-bin interaction without a segment fixed effect. This specification fills segments not previously observed that day using the average speed of the same road type, zone, and time bin;

7. $\mu_{rt} = \alpha_{\ell(r)t}$: a road-type-by-time-bin interaction only;

8. $\mu_{rt} = \alpha_{z(r)\ell(r)}$: a zone-by-road-type interaction without time variation, which assigns remaining cells an average based on location and road class;

9. $\mu_{rt} = \alpha_{\ell(r)}$: a road-type fixed effect only;

10. $\mu_{rt} = \alpha_{z(r)}$: a zone fixed effect only;

11. $\mu_{rt} = \alpha_t$: a time-bin fixed effect only, which assigns the network-wide average speed for each time of day.

We assign the overall daily mean speed to any cells that remain missing after all eleven steps. Finally, we winsorize imputed speeds to the interval $[5, 140]$ km/h. Appendix Figure D5 illustrates the resulting imputed speeds on major roads at one point in time.

Appendix Figure D5: Imputed bus speeds on major roads in the Tel Aviv area, March 4, 2021 (km/h)



The first five specifications include the segment fixed effect ξ_r and therefore use segment-specific information whenever it is available. Within this group, the α terms become progressively coarser: they first omit either zone or road type and then omit time variation. Specifications 6–11 exclude the

segment fixed effect and rely on group averages, allowing us to impute speeds for segments without any same-day observations.

D.3 Imputing Potential Trip Durations and Prices

For each potential habitual trip, departure date, and time, we impute its duration and price if a driver departs on that trip at that time by combining its imputed route constructed in Appendix Section D.1 with the OSM-edge-level bus speeds constructed in Appendix Section D.2.

Segments in the operator data need not align with pricing-zone boundaries. Before imputation, we therefore *collapse* realized trip segments so that a new segment begins only when the pricing zone changes among the Metropolis, Metropolitan fringe, and Periphery. We sum distance, duration, and price within each collapsed segment and identify the segment by its origin and destination TAZs. We then split the OSM route at the entry edge of each new TAZ and assign a subroute to each collapsed segment.

For each trip–date pair, the algorithm traverses the OSM route edge by edge:

1. Decompose the route into consecutive node pairs (u_j, v_j) and obtain the physical length ℓ_j of each edge from the road network.

2. Scale the edge lengths so that the total path length equals the trip's reported distance d:

$$\ell_j^* = \ell_j \cdot \frac{d}{\sum_k \ell_k}.$$

3. For each edge, load the imputed bus speed constructed in Appendix Section D.2, matched by date, edge (u, v), and 15-minute time bin b. Convert the speed from kilometers per hour to meters per second and denote the result by $s_{j,b}$. We set the bus-to-car speed multiplier to 1, so the imputed bus speed serves as the estimated car speed.

4. Traverse edge j at speed $s_{j,b}$, yielding traversal time $\tau_j = \ell_j^*/s_{j,b}$. Cumulative traversal time advances the clock. If edge j crosses a time-bin boundary, split the edge proportionally: apply the current bin's speed to the portion before the boundary and the next bin's speed to the remainder. Repeat this process until all edges have been assigned consistently.

5. Sum traversal times across edges: $\hat{T}_{\text{est}} = \sum_j \tau_j$.

Because location- and trip-specific factors affect the relationship between bus-speed-based estimates and realized driving times, we calibrate the mechanical estimate using a regression on realized trips.

For each realized trip with an imputed route, we first compute $\hat{T}_{\text{est}}$ at the collapsed-segment level using the segment's realized start time and bus speeds on the trip date. We then estimate

$$\log(\text{realized_duration}_{si}) = \beta \cdot \log(\text{estimated_duration}_{si}) + \xi_{\tilde{s}} + \chi_{\tilde{b}_i} + \varepsilon_{si}$$

where s indexes collapsed TAZ-to-TAZ segments, i indexes trips, and b_i denotes the habitual-trip bin for trip i. The terms $\xi_{\tilde{s}}$ and $\chi_{\tilde{b}_i}$ are segment and trip-bin fixed effects, respectively. We pool

segments and bins observed only once into common groups, so that $\tilde{s} = -1$ or $\tilde{b} = -1$ for these singletons.

For each potential trip, whether realized or counterfactual, we exponentiate the fitted value from this regression to obtain the imputed duration:

$$\widehat{T}_{\text{imp}} = \exp\big(\hat{\beta} \cdot \log(\hat{T}_{\text{est}}) + \hat{\xi}_{\tilde{s}} + \hat{\chi}_{\tilde{b}}\big).$$

This transformation calibrates the mechanical travel time multiplicatively.

We compute prices by traversing the same scaled route and accumulating the product of distance and a per-meter tariff that varies with clock time, pricing zone, and the driver's vehicle pollution level. Section 3.1

As in the duration imputation, the algorithm traverses the route edge by edge. At each edge, it uses the collapsed segment's pricing zone, defined by the segment's origin TAZ, determines the pricing period from the current clock time, and accumulates

$$p_j = r_{z,h,v} \cdot \ell_j^*,$$

where $r_{z,h,v}$ is the per-meter rate for zone z, period h, and vehicle pollution level v, and ℓ_j^* is the scaled edge length. If an edge crosses a time-bin boundary, we split the edge and apply the relevant rate to each portion.

We cap the total price at 25 NIS per trip.

For trips with multiple collapsed segments, we advance the start time of each subsequent segment by the cumulative imputed duration of the preceding segments. The pricing period for later segments therefore depends on the regression-calibrated duration of earlier segments. Within each collapsed segment, clock time advances as the trip traverses successive OSM edges, allowing the pricing period to change before the segment ends.

Imputation Accuracy for Realized Trips We assess imputation accuracy using potential-trip observations that are both realized and successfully imputed.

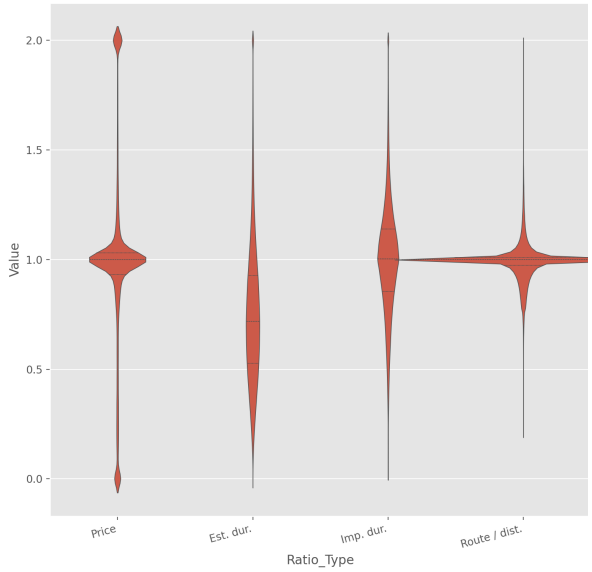
For each outcome, we report the ratio of the modeled value to the realized value; a ratio of 1 indicates equality. *Estimated duration* denotes the output of the bus-speed traversal model before regression calibration, whereas *imputed duration* denotes the output after the fixed-effects regression in logs. The route-length ratio compares the assigned OSM path, scaled to match trip distance, with realized trip distance.

The mechanical duration model systematically understates realized duration: the median ratio of estimated to realized duration is 0.72. The regression-calibrated imputed duration is substantially closer to the realized duration, with a median ratio of 1.00. Thus, although the pre- and post-regression measures use the same route and bus speeds, the regression calibration removes the mechanical estimate's systematic understatement at the median.

The ratio of route length to realized distance is centered near 1. The median imputed-to-realized price ratio is 1.00, with 25th and 75th percentiles of 0.93 and 1.03.

Together, these comparisons show that imputed distances and prices match realized outcomes well near the center of the distribution and that the duration regression corrects a systematically low mechanical travel-time estimate.

Appendix Figure D6: Ratios of imputed to realized trip characteristics



Estimating Fuel Costs for Trips For each potential trip in the demand panel, we construct expected fuel expenditure from the imputed route length.

We use the Israeli government's monthly maximum retail gasoline prices, measured in NIS per liter and inclusive of taxes. These taxes include a 17% value-added tax and an excise tax of approximately 2.6–3.1 NIS per liter during the study period. Appendix Figure D7 plots the resulting gasoline-price series.

We assume homogeneous fuel consumption across trips at an average intensity of 7.9 liters per 100 km, consistent with average fuel-consumption estimates used in external sources.[36]

For potential trip i on date t, let route_length_{it} denote the imputed route length in kilometers. We compute fuel cost as

$$\text{fuel_cost}_{it} = \text{route_length}_{it} \times \bar{\ell} \times p_{m(t)},$$

where $\bar{\ell} = 0.079$ liters per kilometer and $p_{m(t)}$ is the gasoline price in NIS per liter in month $m(t)$.

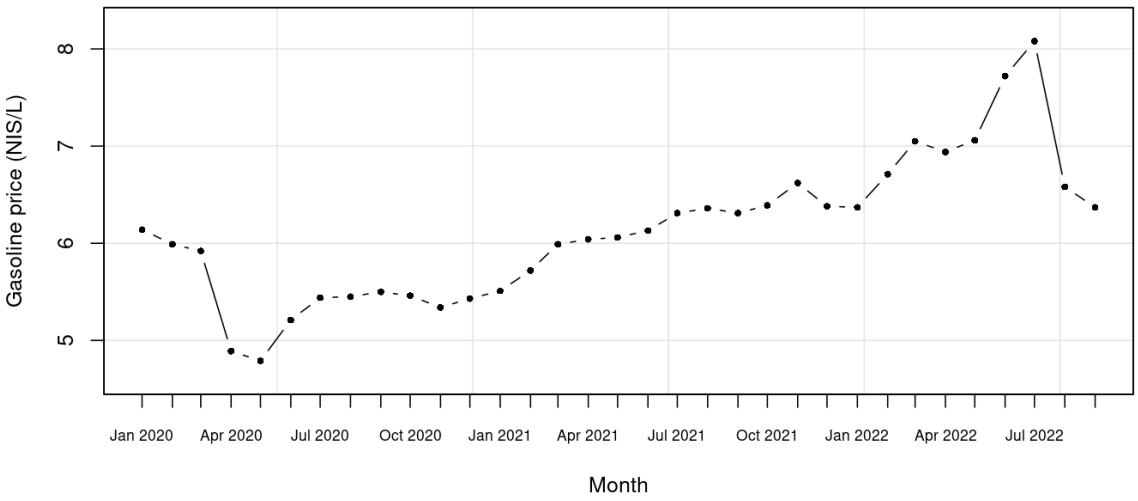
D.4 Duration Instrument Construction

We instrument for travel duration using rainfall intensity along each potential trip's route. This section describes the rainfall data and the construction of the trip-level instrument.

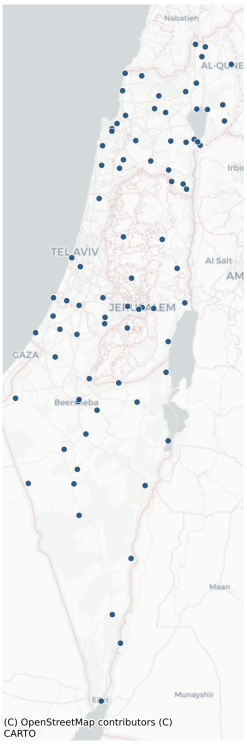
Rainfall Data We use rainfall measured at 10-minute intervals by 84 Israel Meteorological Service (IMS) weather stations. Appendix Figure D8 shows their locations.

[36] For example, see the Global Fuel Economy Initiative: https://www.globalfueleconomy.org/news/event-showcases-argentina-s-new-comparative-vehicle-efficiency-label-and-electric-shift.

Appendix Figure D7: Israeli gasoline prices (NIS per liter, tax-inclusive)

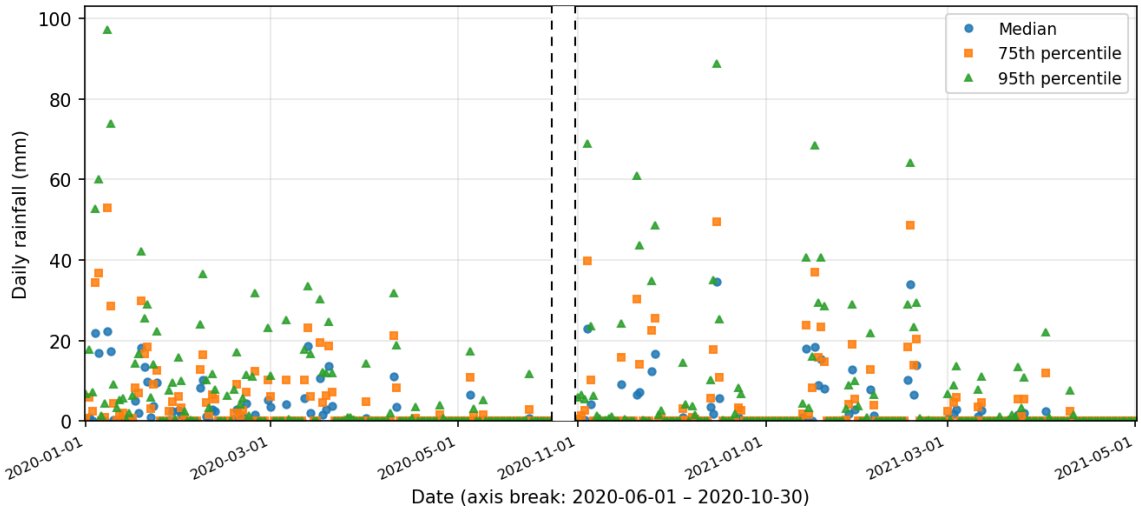


Appendix Figure D8: Locations of weather stations



The weather data cover the sample period from January 2020 through July 2021. Recorded daily rainfall equals 0 mm in approximately 82% of station-day observations. Appendix Figure D9 reports the median and selected quantiles of daily rainfall across stations during rainy periods in the sample.

Appendix Figure D9: Cross-station daily rainfall



IV Construction We begin with the imputed sequence of TAZ polygons traversed by each potential trip. For each TAZ and departure time, we match the nearest weather station with nonmissing rainfall data. If rainfall is missing at all three nearest stations, we set the instrument to missing.

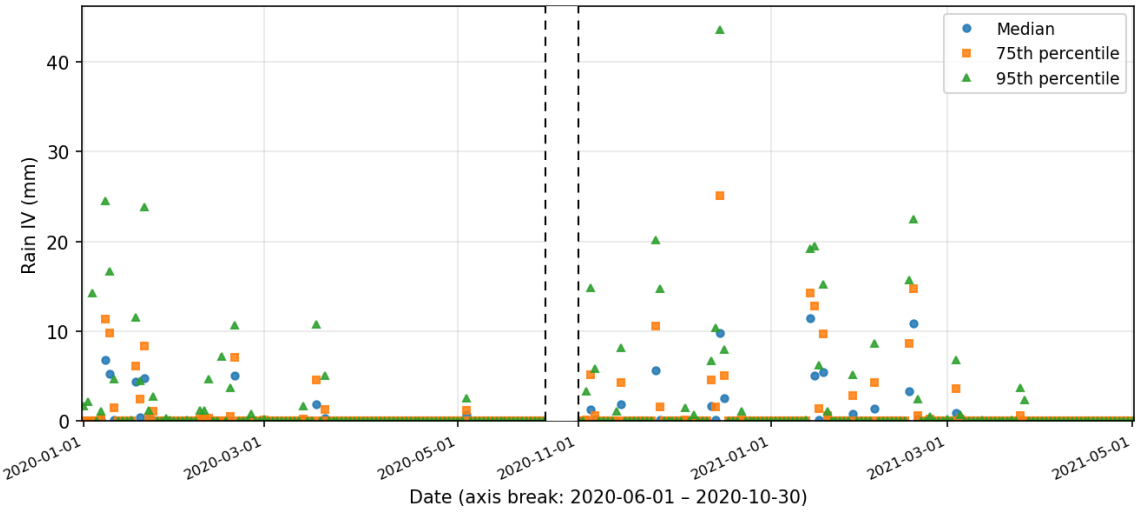
For each TAZ along the route, we then compute cumulative precipitation at the matched station from 1 PM through the trip's departure time.

We define trip-level exposure to afternoon rainfall as the distance-weighted average of cumulative precipitation across the TAZs traversed by the route:

$$\texttt{trip_iv} = \frac{\sum_{r \in \text{route}} w_r \cdot \texttt{taz_rain}_r}{\sum_{r \in \text{route}} w_r},$$

where w_r is the estimated distance traveled within TAZ r, and $\texttt{taz_rain}_r$ is cumulative precipitation from 1 PM through the trip's departure time at the station matched to TAZ r. Appendix Figure D10 shows the daily cross-sectional distribution of the resulting instrument.

Appendix Figure D10: Daily cross-section of trip IV values



D.5 Public Transit Alternatives for Habitual Trips

For each origin–destination pair and relevant departure time in the habitual-trip sample, we construct a feasible public transit itinerary, its expected travel time, and its modal composition. We construct these alternatives with OpenTripPlanner (OTP), an open-source routing engine for multimodal travel itineraries.

Input Data and OTP Version OTP requires an origin coordinate, a destination coordinate, and a desired departure time. For each potential trip, we assign the centroids of the origin and destination TAZs as coordinates. We query the habitual trip's median departure time on February 21, 2021, a date near the midpoint of the sample that reflects stable post-COVID-19 transit service. We use OTP version 2.1.0 for routing.[37] The routing graph combines a January 2021 OpenStreetMap extract for the road and pedestrian network with the Israeli Ministry of Transport's February 2021 General Transit Feed Specification (GTFS) schedule for the public transit network.

OTP may return multiple candidate itineraries for a single origin–destination–time query. For each successful query, we select the valid itinerary with the shortest total duration, rather than the first itinerary returned or the itinerary with the earliest arrival time. Alternative selection rules may therefore choose different itineraries for some queries.

Travel-Time Components We decompose the selected itinerary into walking time, waiting time, in-vehicle transit time, and total duration. Total duration runs from the itinerary's start time to its end time. Walking time sums the access, transfer, and egress walking legs, while transit time sums all in-vehicle public transit legs. Where applicable, we use the timing information returned by OTP to compute waiting time as the residual between total duration and the other components.

[37] See the OTP 2.1.0 documentation: https://docs.opentripplanner.org/en/v2.1.0/.

Mode Classification We classify each selected itinerary by the modes used in its transit legs, as reported in the OTP response. From these fields, we construct indicators for bus-only, rail-including, and walk-only itineraries. For train itineraries, the final fare combines the matched train fare and the bus fare, subject to the applicable fare cap. We calculate the fare cap for train-including itineraries from `air_distance_km`. For itineraries without a train leg, we leave the train-fare fields missing or set them to zero. The Israeli pricing system applies different fare caps to train and non-train trips.

Modeling the Counterfactual Tel Aviv Metro To model the planned Tel Aviv metro system, we construct a second OTP network that adds the planned metro service to the baseline network. Comparing the two networks identifies trips for which the metro changes the fastest public transit route, total travel time, or modal composition.

We construct the metro network by merging the baseline February 2021 GTFS feed with the planned metro GTFS feed. The merged feed preserves the baseline public transit network and adds metro routes, stops, stop times, trips, calendars, and fare information. We adjust the metro calendar so that the added service operates during the query period used for the baseline OTP runs.

We construct the metro GTFS from version 0.33 of the national transportation model maintained by Netivei Israel.[38] The model contains forecasts for transportation services that do not yet exist but are expected to operate in the future, including the metro system.

After merging the GTFS feeds, we rebuild the OTP graph using the same OpenStreetMap extract and OTP version as in the baseline procedure. The addition of metro service to the public transit schedule is therefore the only difference between the routing environments. We query the metro OTP server using the same origin–destination–time rows as in the baseline construction. The metro and baseline runs also use the same itinerary-selection rule: for each query, we select the valid itinerary with the shortest total duration.

E Demand Estimation Details

In this appendix, we describe the procedure used to estimate the trip-demand model in Section 6.1. The estimator combines the panel likelihood of drivers' repeated departure-time choices for their habitual trips with an IV-GMM penalty on market-level mean utility, following the logic of Grieco et al. (2025). We first define the likelihood and the market-level decomposition, then describe the simulated-likelihood estimation and inference algorithms.

[38] A complete explanation, available only in Hebrew, is provided at `https://tinyurl.com/3sjuc7kv`.

E.1 Estimator

For each potential trip departure time (i, ℓ, t), collect expected duration $\hat{D}_{i\ell t}$, price $\hat{P}_{i\ell t}$, and the driving-to-public-transit-time ratio R_i in

$$x_{i\ell t} \coloneqq \begin{pmatrix} \hat{D}_{i\ell t} \\ \hat{P}_{i\ell t} \\ R_i \end{pmatrix}.$$

We parameterize the driver-specific coefficient vector as

$$\beta(X_i, \nu_i) \coloneqq \begin{pmatrix} \gamma(X_i, \nu_i) \\ \alpha(X_i, \nu_i) \\ \rho(X_i, \nu_i) \end{pmatrix} \coloneqq \underbrace{\begin{pmatrix} \gamma_0 \\ \alpha_0 \\ \rho_0 \end{pmatrix}}_{\beta_0} + \Pi X_i + L\nu_i, \qquad \nu_i \sim \mathcal{N}(0, I_3). \tag{18}$$

Here, γ, α, and ρ are the coefficients on duration, price, and the public-transit ratio. The vector X_i contains standardized socio-economic status and population density for the driver's home neighborhood. The matrix Π contains their interactions with the three trip characteristics, and $\Sigma = LL'$ is a general 3×3 random-coefficient covariance matrix parameterized by the lower-triangular Cholesky factor L. For the optimization below, it is convenient to collect all fixed coefficients in

$$b \coloneqq \begin{pmatrix} \beta_0 \\ \mathrm{vec}(\Pi) \end{pmatrix}.$$

We estimate the model using a nested objective that jointly optimizes the market-level utilities and fixed coefficients for each candidate value of L:

$$\begin{aligned} \hat{L} &\in \arg\min_{L} \mathcal{Q}(L), \\ \mathcal{Q}(L) &\coloneqq \min_{\delta, b} \mathcal{L}(\delta, b, L), \\ \mathcal{L}(\delta, b, L) &\coloneqq -\frac{1}{N} \sum_{i=1}^{N} \log p_i(\delta, b, L) + \frac{1}{N} \mathcal{P}(\delta, \beta_0). \end{aligned} \tag{19}$$

The vector δ collects market-by-departure-time mean utilities, $p_i(\delta, b, L)$ is driver i's panel marginal likelihood of their sequence of trip departure choices after integrating over unobserved heterogeneity ν, and $\mathcal{P}(\delta, \beta_0)$ is the IV-GMM penalty, all defined below. At the solution, $(\hat{\delta}, \hat{b})$ are the minimizers of the inner problem evaluated at $\hat{L}$.

Market-Level Utility Decomposition For each market-period (m, t), define

$$\overline{x}_{mt} \coloneqq \begin{pmatrix} \overline{D}_{mt} \\ 0 \\ \overline{R}_{mt} \end{pmatrix},$$

where $\overline{D}_{mt}$ and $\overline{R}_{mt}$ average $\hat{D}_{i\ell t}$ and R_i across the trip opportunities in market-period (m,t). The zero in the price position is a modeling restriction: market-mean price does not enter mean utility. We instead identify the price coefficient from the experimental price variation within drivers' panels. Market-level utility decomposes as

$$\begin{aligned}\delta_{mt} &= \hat{V}_{mt} + \overline{x}'_{mt}\beta_0 = \hat{V}_{mt} + \gamma_0\overline{D}_{mt} + \rho_0\overline{R}_{mt},\\ \hat{V}_{mt} &= \delta_{mt} - \overline{x}'_{mt}\beta_0.\end{aligned} \tag{20}$$

The residual $\hat{V}_{mt}$ captures scheduling preferences and other unobserved determinants of demand for departing in period t in market m.

Stacking $\overline{x}'_{mt}$ as the rows of $\overline{X}$ gives $\hat{V} = \delta - \overline{X}\beta_0$. We construct a more powerful instrument for market-mean duration following the partially linear optimal-instrument approach of Chen et al. (2021). Within each departure-time option, we use random forests to approximate the required conditional expectations, using out-of-bag predictions to limit overfitting (Athey et al., 2019). The excluded predictor is the route-level afternoon-rainfall measure described in Appendix Section D, while the market-mean transit ratio enters the prediction linearly as an exogenous characteristic. We collect resulting optimal duration instrument and the market-mean transit ratio, which serves as its own instrument, into the matrix Z whose rows correspond to market-periods. Define the instrument projection matrix $P_Z := Z(Z'Z)^{-1}Z'$; then the market-level decomposition induces the IV-GMM penalty

$$\mathcal{P}(\delta,\beta_0) := \tfrac{1}{2}(\delta - \overline{X}\beta_0)'P_Z(\delta - \overline{X}\beta_0) = \tfrac{1}{2}\hat{V}'P_Z\hat{V}. \tag{21}$$

Thus, the penalty disciplines the mean duration and transit-ratio coefficients through market-level moments, while the price column of $\overline{X}$ is zero by construction.

Driver-Specific Marginal Likelihood Conditional on a draw of ν, write the systematic utility of departure time t on date ℓ as

$$\begin{aligned}u_{i\ell t}(\delta,b,L,\nu) := \delta_{m(i,\ell)t} + \beta_0'\left(x_{i\ell t} - \overline{x}_{m(i,\ell)t}\right)\\ + \left(\beta(X_i,\nu) - \beta_0\right)' x_{i\ell t},\end{aligned} \tag{22}$$

with outside-option utility normalized to $u_{i\ell\infty} = 0$. Substituting Equation (20) into Equation (22) yields the equivalent representation

$$u_{i\ell t}(\delta,b,L,\nu) = \hat{V}_{m(i,\ell)t} + \beta(X_i,\nu)'x_{i\ell t}.$$

The corresponding conditional choice probabilities are

$$s_{i\ell t}(\delta,b,L,\nu) := \frac{\exp(u_{i\ell t}(\delta,b,L,\nu))}{1 + \sum_{t'\in\mathcal{T}_i}\exp(u_{i\ell t'}(\delta,b,L,\nu))},$$

with outside-option probability

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

Let $Y_{i\ell\infty} \coloneqq 1 - \sum_{t \in \mathcal{T}_i} Y_{i\ell t}$ indicate that driver i chose the outside option on date ℓ. The driver's marginal likelihood contribution is

$$p_i(\delta, b, L) \coloneqq \int \prod_{\ell \in \mathcal{L}_i} \left[\prod_{t \in \mathcal{T}_i} s_{i\ell t}(\delta, b, L, \nu)^{Y_{i\ell t}} \cdot s_{i\ell\infty}(\delta, b, L, \nu)^{Y_{i\ell\infty}} \right] \phi_3(\nu)\, d\nu, \tag{23}$$

where ϕ_3 is the standard trivariate normal density. Integrating a common ν_i over the entire sequence of choices preserves the panel structure: repeated choices by the same driver update the likelihood of that driver's duration, price, and transit-ratio tastes jointly.

E.2 Simulated Likelihood and Optimization

Since the integral in Equation (23) has no closed form, we evaluate it using a numerically convenient importance sampling algorithm, and we optimize the objective function using a nested procedure that jointly estimates (δ, b) for each candidate random-coefficient Cholesky factor L.

Importance Sampling for Likelihood Evaluation For a given driver, the integrand is the prior density over tastes multiplied by the likelihood of the driver's entire panel of choices. Because the panel likelihood multiplies many date-level probabilities, values of ν that fit the observed sequence even modestly better receive much more weight than values that fit it poorly. The integrand therefore becomes sharply concentrated once a driver contributes many repeated choices. Drawing directly from the diffuse standard-normal distribution ϕ_3 would yield simulation draws mostly in regions where the integrand is effectively zero, causing numerical instability.

Instead, we use an importance sampling procedure with driver-specific proposal distributions that concentrates draws near the posterior mass of ν given each driver's observed choices. For fixed (δ, b, L), define driver i's conditional log-likelihood as

$$\ell_i(\nu) \coloneqq \sum_{\ell \in \mathcal{L}_i} \left[\sum_{t \in \mathcal{T}_i} Y_{i\ell t} \log s_{i\ell t}(\delta, b, L, \nu) + Y_{i\ell\infty} \log s_{i\ell\infty}(\delta, b, L, \nu) \right]. \tag{24}$$

The integrand in Equation (23) is proportional to

$$\exp\left(\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2\right),$$

which is the posterior kernel for ν_i up to a normalizing constant (see Revelt and Train (2000) and Chapter 11 in Train (2009)). We use a Laplace approximation to this posterior distribution as our importance-sampling proposal, which is most accurate when the posterior is unimodal and concentrated (precisely when numerical instability is most likely to occur).

Specifically, let $\hat{\nu}_i$ denote the posterior mode,

$$\hat{\nu}_i \coloneqq \arg\max_{\nu} \left[\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2\right]. \tag{25}$$

A second-order 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 H_i^{-1} (see Chapter 11 in Train (2009)). The posterior mode and Hessian are inexpensive to compute via Newton's method and separable across drivers. At each Newton step, the posterior gradient and negative Hessian are

$$\begin{aligned} \nabla_\nu \left[\ell_i(\nu) - \tfrac{1}{2}\|\nu\|^2\right] &= L' \frac{\partial \ell_i(\nu)}{\partial \beta_i} - \nu, \\ H_i(\nu) &= I_3 - L' \frac{\partial^2 \ell_i(\nu)}{\partial \beta_i \partial \beta_i'} L, \end{aligned} \tag{26}$$

where $\beta_i = \beta_0 + \Pi X_i + L\nu$. At the mode, write

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

for its Cholesky factorization.

Given these Normal approximation parameterizations for the driver-specific posteriors, we draw common base vectors $\{z_r\}_{r=1}^R$ from $\mathcal{N}(0, I_3)$ and form the driver-specific proposal draws

$$\nu_{ir} \coloneqq \hat{\nu}_i + U_i^{-1} z_r, \qquad r = 1, \ldots, R. \tag{27}$$

Each draw receives log importance weight (computed in logs for numerical stability with small weights)

$$\log w_{ir} \coloneqq \ell_i(\nu_{ir}) - \tfrac{1}{2}\|\nu_{ir}\|^2 + \log|\det U_i^{-1}| + \tfrac{1}{2}\|z_r\|^2. \tag{28}$$

The first two terms give the log of the unnormalized posterior density at the draw; the final two divide by the Gaussian proposal density induced by Equation (27). The normalized weights are

$$a_{ir} \coloneqq \frac{\exp(\log w_{ir})}{\sum_{r'=1}^R \exp(\log w_{ir'})}. \tag{29}$$

We approximate driver i's log marginal likelihood by

$$\log \hat{p}_i = \log\left(\sum_{r=1}^R e^{\log w_{ir}}\right) - \log R, \tag{30}$$

which we evaluate using the log-sum-exp transformation.

Computing Gradients Within an optimization iteration, we hold fixed the proposal draws in Equation (27) and recompute their weights using Equation (29) at every candidate parameter value. Differentiating Equation (30) then gives the posterior-average-score identity

$$\frac{\partial \log \hat{p}_i}{\partial \psi} = \sum_{r=1}^{R} a_{ir} \frac{\partial \ell_i(\nu_{ir})}{\partial \psi} \tag{31}$$

for any parameter ψ entering $u_{i\ell t}(\delta, b, L, \nu)$. This identity converts each simulated-likelihood gradient into a posterior-weighted average of conditional logit scores.

For δ_{mt}, the draw-specific score is

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

Consequently,

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

The objective gradient with respect to the market-level utilities is therefore

$$\nabla_\delta \mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \nabla_\delta \log \hat{p}_i + \frac{1}{N} P_Z \hat{V}. \tag{32}$$

To write the gradient for b compactly, define the utility derivative

$$q_{i\ell t} \coloneqq \begin{pmatrix} x_{i\ell t} - \overline{x}_{m(i,\ell)t} \\ X_i \otimes x_{i\ell t} \end{pmatrix}, \tag{33}$$

where the first three entries correspond to β_0 and the remaining entries correspond to vec(Π). Then

$$\frac{\partial \log \hat{p}_i}{\partial b} = \sum_{\ell \in \mathcal{L}_i} \sum_{t \in \mathcal{T}_i} \sum_{r=1}^{R} a_{ir} \left[Y_{i\ell t} - s_{i\ell t}(\delta, b, L, \nu_{ir}) \right] q_{i\ell t}. \tag{34}$$

Let S_β be the matrix that selects β_0 from b, so that $\beta_0 = S_\beta b$. Since the penalty depends on b only through β_0, its fixed-coefficient gradient is $-S_\beta' \overline{X}' P_Z \hat{V}$. Combining the likelihood and penalty terms gives

$$\nabla_b \mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \nabla_b \log \hat{p}_i - \frac{1}{N} S_\beta' \overline{X}' P_Z \hat{V}. \tag{35}$$

Because the price column of $\overline{X}$ is zero, the penalty component of the price-coefficient score is zero. The same is true for the demographic interactions, which enter individual utility but not the market-level decomposition.

Finally, changing L_{kq} shifts the coefficient on characteristic k in the direction $\nu_{ir,q}$. The corre-

sponding likelihood gradient is

$$\frac{\partial \log \hat{p}_i}{\partial L_{kq}} = \sum_{r=1}^{R} a_{ir} \sum_{\ell \in \mathcal{L}_i} \sum_{t \in \mathcal{T}_i} \left[Y_{i\ell t} - s_{i\ell t}(\delta, b, L, \nu_{ir}) \right] x_{i\ell t,k} \nu_{ir,q}, \tag{36}$$

for the free lower-triangular elements of L. Since the penalty does not depend directly on L, the outer gradient is the negative average of these likelihood scores evaluated at the inner solution. By the envelope theorem,

$$\nabla_{\mathrm{vech}(L)} \mathcal{Q}(L) = \nabla_{\mathrm{vech}(L)} \mathcal{L}(\delta, b, L)\big|_{(\delta,b)=(\hat{\delta}(L),\hat{b}(L))}. \tag{37}$$

Optimization Algorithm We compute the estimator with an inner L-BFGS problem over (δ, b) and an outer L-BFGS problem over the free elements of L. This partition matches the estimator in Equation (19): the IV-GMM penalty and all fixed taste coefficients are handled jointly in the inner problem, rather than through a separate closed-form IV update.

1. **Initialize.** Choose starting values for (δ, b, L) and choose the number R of simulation draws. Set convergence tolerances for the Newton, inner L-BFGS, and outer L-BFGS iterations. Draw

   $$z_r \sim \mathcal{N}(0, I_3), \qquad r = 1, \ldots, R,$$

   once and hold the base draws fixed throughout estimation. These common random numbers stabilize comparisons of the simulated objective across parameter values.

2. **Compute Laplace proposals.** At outer iteration k, fix $L = L^{(k)}$ and use the current, warm-started values of (δ, b) to solve Equation (25) for each driver by damped Newton iterations. Starting from the preceding outer iteration's mode, or from zero at initialization, iterate

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

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

   where $\lambda_i^{(k,n)}$ is chosen by backtracking. The driver-specific mode problems are separable and can be solved in parallel. At convergence, set

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

3. **Generate proposal draws.** Form

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

   Hold these proposal draws and their proposal-density terms fixed during the inner solve. For every candidate (δ, b), recompute the target likelihood, importance weights, normalized weights, and $\log \hat{p}_i$ using Equation (28), Equation (29), and Equation (30).

4. **Joint inner solve for (δ, b).** Holding $L^{(k)}$ fixed, solve

$$(\hat{\delta}^{(k)}, \hat{b}^{(k)}) \in \arg\min_{\delta, b} \mathcal{L}(\delta, b, L^{(k)})$$

by L-BFGS using the analytic gradients in Equation (32) and Equation (35). This step estimates the market-level utilities, all three mean taste coefficients, and the demographic interactions jointly. Because the distribution of random coefficients is Gaussian, $\mathcal{L}$ is jointly convex in (δ, b).[39]

5. **Evaluate the profiled objective and gradient in L.** Set

$$\mathcal{Q}(L^{(k)}) = \mathcal{L}(\hat{\delta}^{(k)}, \hat{b}^{(k)}, L^{(k)})$$

and evaluate its gradient from Equation (36) and Equation (37).

6. **Outer L-BFGS step.** Take an outer step in the free elements of L to obtain $L^{(k+1)}$. Because changing L changes the posterior curvature and the Laplace proposal, recompute the posterior modes, Hessians, proposal draws, and importance weights at the next outer iteration. Warm-start the next inner solve from $(\hat{\delta}^{(k)}, \hat{b}^{(k)})$.

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

At convergence, the estimator combines the jointly optimized $(\hat{\delta}, \hat{\beta}_0, \hat{\Pi})$ with the profiled random-coefficient factor $\hat{L}$.

E.3 Inference

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

$$\theta := \begin{pmatrix} b \\ \operatorname{vech}(L) \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \operatorname{vec}(\Pi) \\ \operatorname{vech}(L) \end{pmatrix},$$

treating the high-dimensional market-level utility vector δ as a nuisance parameter. This parameter vector includes the mean transit-ratio coefficient, its demographic interactions, and all six distinct elements of the 3×3 random-coefficient covariance factor.

Let H denote the full Hessian of $\mathcal{L}$ with respect to (δ, θ). Profiling out δ gives the Schur-complement Hessian

$$\widetilde{H}_{\theta\theta} := H_{\theta\theta} - H_{\theta\delta} H_{\delta\delta}^{-1} H_{\delta\theta}. \tag{38}$$

[39] Conditional on any ν, systematic utility $u_{i\ell t}$ is affine in (δ, b) and log-concavity is closed under multiplication, so the conditional logit likelihood is log-concave in (δ, b, ν). Multiplying by the log-concave Gaussian density of ν preserves joint log-concavity, as does integrating over ν (Boyd and Vandenberghe, 2004, p. 106). The negative marginal log-likelihood is therefore convex, as is its population expectation. The IV-GMM penalty is also convex because it is a positive-semidefinite quadratic form in $\delta - \overline{X}\beta_0$. A finite fixed-draw importance-sampling approximation need not inherit global convexity exactly: the logarithm of a finite weighted sum of conditional likelihoods need not be concave for every realization of the draws. However, as the number of well-targeted draws increases, the simulated objective approaches the exact objective and is therefore approximately convex in practice.

The computational objective profiles out (δ, b) while optimizing over L. For inference, the Schur complement in Equation (38) gives the curvature for the low-dimensional taste parameters after allowing the market-level utilities δ to adjust.

The matrix $H_{\delta\delta}$ is large but structured. Stack the mean utilities as

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

where $\mathcal{T}_m$ is the set of departure periods available in market m. For each driver i, draw r, and market m, define

$$s_{i\ell,m}^{(r)} \coloneqq \big(s_{i\ell t}(\delta, b, L, \nu_{ir})\big)_{t\in\mathcal{T}_m}, \qquad Y_{i\ell,m} \coloneqq (Y_{i\ell t})_{t\in\mathcal{T}_m},$$

and the draw-specific score and curvature contributions

$$\begin{aligned} G_{imr} &\coloneqq \sum_{\ell:\, m(i,\ell)=m} \left(Y_{i\ell,m} - s_{i\ell,m}^{(r)}\right), \\ \Lambda_{imr} &\coloneqq \sum_{\ell:\, m(i,\ell)=m} \left[\operatorname{diag}(s_{i\ell,m}^{(r)}) - s_{i\ell,m}^{(r)} s_{i\ell,m}^{(r)\prime}\right]. \end{aligned}$$

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 the posterior-average score in Equation (31) once more gives

$$\begin{aligned} \frac{\partial^2 \log \hat{p}_i}{\partial \delta_m \partial \delta_m'} &= -\sum_{r=1}^{R} a_{ir}\Lambda_{imr} + \sum_{r=1}^{R} a_{ir}(G_{imr} - \overline{G}_{im})(G_{imr} - \overline{G}_{im})', \\ \overline{G}_{im} &\coloneqq \sum_{r=1}^{R} a_{ir} G_{imr}. \end{aligned}$$

Accordingly, define the market-m likelihood-curvature block as

$$B_m \coloneqq -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \log \hat{p}_i}{\partial \delta_m \partial \delta_m'}, \qquad B \coloneqq \operatorname{diag}(B_1, \ldots, B_{|\mathcal{M}|}).$$

Each block averages the simulated-likelihood curvature with respect to the departure-time utilities in one market.

The penalty contributes the low-rank matrix $N^{-1}P_Z$. Setting

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

the nuisance-parameter Hessian is

$$H_{\delta\delta} = B + ZC_ZZ'.$$

We evaluate its inverse using the Woodbury identity

$$H_{\delta\delta}^{-1} = B^{-1} - B^{-1}Z\left(C_Z^{-1} + Z'B^{-1}Z\right)^{-1}Z'B^{-1}, \tag{39}$$

which avoids forming a dense inverse: it just requires independent market-level solves involving B_m and one low-dimensional solve whose dimension equals the number of instruments (Grieco et al., 2025).

To construct the meat of the sandwich covariance, we profile both the likelihood and penalty scores with respect to δ. For driver i, define the likelihood-score contributions

$$g_{i,\delta} \coloneqq -\frac{1}{N}\nabla_\delta \log \hat{p}_i, \qquad g_{i,\theta} \coloneqq -\frac{1}{N}\begin{pmatrix} \nabla_b \log \hat{p}_i \\ \nabla_{\mathrm{vech}(L)} \log \hat{p}_i \end{pmatrix}. \tag{40}$$

The first block of $g_{i,\theta}$ is given by Equation (34), and the second is given by Equation (36). In particular, the score for the three mean coefficients is

$$\frac{\partial \log \hat{p}_i}{\partial \beta_0} = \sum_{\ell\in\mathcal{L}_i}\sum_{t\in\mathcal{T}_i}\sum_{r=1}^{R} a_{ir}\left[Y_{i\ell t} - s_{i\ell t}(\delta, b, L, \nu_{ir})\right]\left(x_{i\ell t} - \overline{x}_{m(i,\ell)t}\right).$$

The duration and transit-ratio entries use deviations from their market-period means, whereas the price entry uses price in levels because the price component of $\overline{x}_{mt}$ is zero.

The corresponding profiled driver score is

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

For the penalty scores, let $c = 1, \ldots, n_{\mathrm{mp}}$ index the market-period rows of $(\delta, \overline{X}, Z)$, let e_c denote the corresponding standard basis vector in the δ-space, and let $\overline{x}_c$ denote the cth row of $\overline{X}$ written as a column. Define

$$\begin{aligned} g_{c,\delta}^{\mathrm{pen}} &\coloneqq \frac{1}{N}e_c(P_Z\hat{V})_c, \\ g_{c,\theta}^{\mathrm{pen}} &\coloneqq -\frac{1}{N}\begin{pmatrix} S_\beta'\overline{x}_c(P_Z\hat{V})_c \\ 0_{\dim(\mathrm{vech}(L))} \end{pmatrix}. \end{aligned} \tag{41}$$

Summing Equation (41) over c recovers the full penalty gradients in Equation (32) and Equation (35). Because $\overline{x}_c = (\overline{D}_c, 0, \overline{R}_c)'$, the penalty contributes to the duration and transit-ratio scores. It does not contribute to the price score, the scores on demographic interactions, or the random-coefficient scores. Profile each penalty contribution as

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

The robust sandwich covariance estimator is

$$\widehat{V}_\theta = \widetilde{H}_{\theta\theta}^{-1}\left(\sum_{i=1}^{N}\widetilde{g}_{i,\theta}\widetilde{g}_{i,\theta}' + \sum_{c=1}^{n_{\mathrm{mp}}}\widetilde{g}_{c,\theta}^{\mathrm{pen}}\left(\widetilde{g}_{c,\theta}^{\mathrm{pen}}\right)'\right)\widetilde{H}_{\theta\theta}^{-1}. \tag{42}$$

This covariance allows for misspecification of the random-coefficient distribution and for the sampling

variation induced by the market-level moment penalty.

Although we optimize over L, the covariance matrix $\Sigma = LL'$ is more interpretable. We therefore use the delta method for the reported parameterization

$$\vartheta := \begin{pmatrix} b \\ \operatorname{vech}(\Sigma) \end{pmatrix}.$$

If

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

then

$$\widehat{V}_\vartheta = J \widehat{V}_\theta J'. \tag{43}$$

We use Equation (42) and Equation (43) to report the standard errors and confidence intervals in Appendix Table 1.

E.4 The Distribution of Preferences Across Drivers

Having estimated the demand parameters, we use the same simulation objects to construct the distribution of driver-level duration, price, and transit-ratio coefficients. We use the posterior distribution rather than the unconditional preference distribution because each driver's repeated choices reveal information about that driver's tastes given on the estimated population preference distribution (see Revelt and Train (2000) and Chapter 11 in Train (2009)).

Let $Y_i := \{Y_{i\ell t}\}_{\ell,t}$ denote the panel of choices observed for driver i, and let $\hat{\theta}_{\text{full}}$ collect $(\hat{\delta}, \hat{\beta}_0, \hat{\Pi}, \hat{L})$. At these estimates, the posterior density of driver i's unobserved taste draw is proportional to

$$p_i(\nu \mid Y_i; \hat{\theta}_{\text{full}}) \propto \exp\Big(\hat{\ell}_i(\nu)\Big)\, \phi_3(\nu),$$

where $\hat{\ell}_i(\nu)$ is the conditional log-likelihood in Equation (24) evaluated at $\hat{\theta}_{\text{full}}$. We note that $p_i(\nu \mid Y_i; \hat{\theta}_{\text{full}})$ is the same posterior kernel used for importance sampling.

Instead of drawing a new set of random numbers, we reuse the fixed three-dimensional base draws $\{z_r\}_{r=1}^R$ used in estimation. With posterior mode $\hat{\nu}_i$ and Hessian factor $\hat{U}_i$ evaluated at $\hat{\theta}_{\text{full}}$, form

$$\widetilde{\nu}_{ir} := \hat{\nu}_i + \hat{U}_i^{-1} z_r, \qquad r = 1, \ldots, R,$$

and assign each draw normalized importance weight

$$\hat{a}_{ir} := \frac{\exp(\log \hat{w}_{ir})}{\sum_{r'=1}^R \exp(\log \hat{w}_{ir'})},$$

where $\log \hat{w}_{ir}$ is Equation (28) evaluated at $\hat{\theta}_{\text{full}}$ and $\widetilde{\nu}_{ir}$. These weights convert the Laplace-proposal draws into a discrete approximation to $p_i(\nu \mid Y_i; \hat{\theta}_{\text{full}})$.

The corresponding random-coefficient draw for driver i is

$$\hat{\beta}_{ir} \coloneqq \begin{pmatrix} \hat{\gamma}_{ir} \\ \hat{\alpha}_{ir} \\ \hat{\rho}_{ir} \end{pmatrix} = \begin{pmatrix} \hat{\gamma}_0 \\ \hat{\alpha}_0 \\ \hat{\rho}_0 \end{pmatrix} + \hat{\Pi} X_i + \hat{L}\tilde{\nu}_{ir}.$$

Thus, driver i's posterior preference distribution over $\mathbb{R}^3$ is the weighted empirical distribution

$$\hat{F}_i(A) = \sum_{r=1}^{R} \hat{a}_{ir} 1\{\hat{\beta}_{ir} \in A\},$$

for any measurable set $A \subseteq \mathbb{R}^3$. When reporting the distribution across drivers, we pool the driver-specific posterior draws using weights ω_i:

$$\hat{F}(A) = \sum_{i=1}^{N} \omega_i \sum_{r=1}^{R} \hat{a}_{ir} 1\{\hat{\beta}_{ir} \in A\},$$

where $\sum_i \omega_i = 1$ and $\omega_i = 1/N$ unless we explicitly reweight drivers to match a target population, e.g. in Appendix Section F.3. The estimated value-of-time distribution plotted in Appendix Figure A7 is obtained from the posterior ratio $\hat{\gamma}_{ir}/\hat{\alpha}_{ir}$.

F Traffic Flow Supply Estimation Details

This appendix provides estimation and implementation details for the supply-side inputs summarized in Section 6.2.

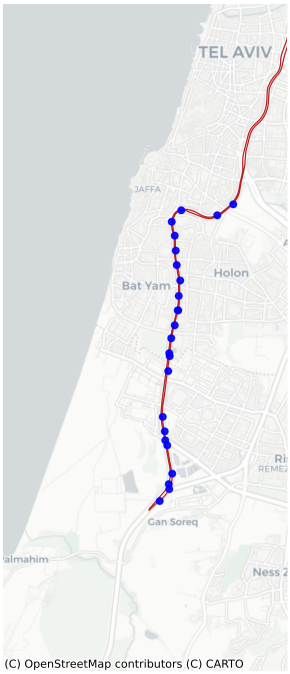
F.1 Segment-Level Congestion Function Estimation

We construct the estimation sample from available sensor cells recorded between 2020 and 2022. Each cell reports total traffic volume and median speed for a given lane over a 5-minute interval. For each observation, we compute lane-level density (in vehicles per kilometer) by dividing lane-level traffic volume by median speed (in kilometers per hour); equivalently, lane-level flow is the product of density and median speed. The final estimation sample contains about 1.2 million observations from 24 sensors, ranging from roughly 15,000 to 117,000 observations per sensor, with a median of about 47,000. Appendix Figure F1 shows the locations of the sensors used for estimation.

Next, we group the sensors into classes based on their OSM-derived speed limits (70 or 90 km/h) and road types (either regular roads or tunnels/bridges), yielding four baseline groups. Within each group, we estimate a lane-level congestion function that satisfies three shape restrictions: (1) the flow function is concave in density; (2) flow equals zero at zero density; and (3) the slope at zero density equals a free-flow speed proxy based on the group's maximum allowed speed.

In particular, let ρ denote lane-level density (vehicles per km) and let $f(\rho) = \rho v(\rho)$ denote flow (vehicles per hour), where $v(\rho)$ is median speed in the data. We approximate f using a Bernstein

Appendix Figure F1: Ayalon and estimation-sample APEG locations



polynomial of degree 20:

$$f(\rho; b) \coloneqq \sum_{k=0}^{20} b_k B_k^{20}(x), \qquad x = T(\rho) \in [0, 1],$$

where b_k are coefficients, $T(\rho)$ is an affine rescaling of observed density into the unit interval, and the Bernstein basis functions of degree n are given by

$$B_k^n(x) = \binom{n}{k} x^k (1-x)^{n-k}, \qquad k = 0, \ldots, n.$$

We estimate the coefficients b_k using linearly-constrained least squares, where $F_i = \rho_i V_i$ and V_i denote the observed flow and median speed for a given sensor observation i:

$$\hat{b} \coloneqq \underset{b \in \mathbb{R}^{21}}{\arg\min} \quad \sum_{i=1}^{N} (F_i - f(\rho_i; b))^2 \tag{44}$$

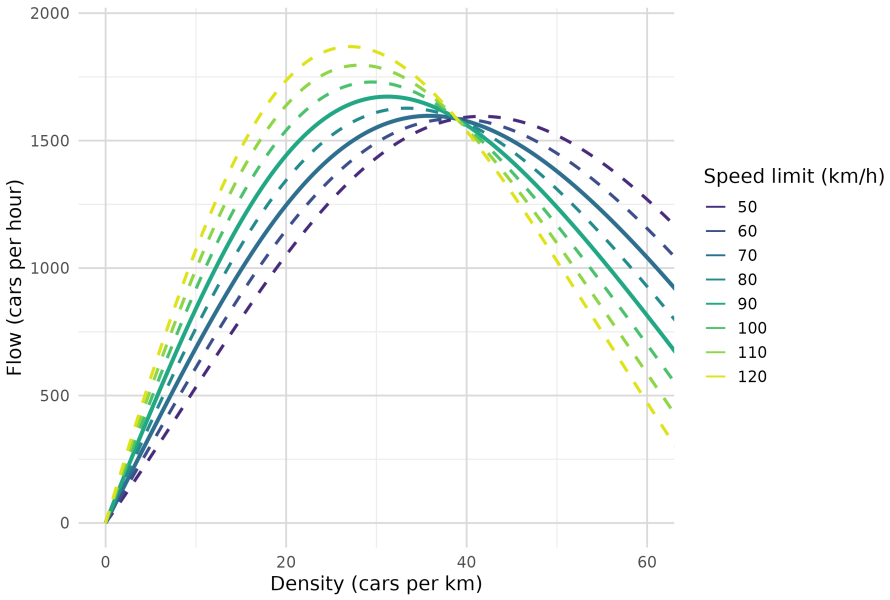
$$\text{subject to} \quad b_{k+1} - 2b_k + b_{k-1} \leq 0, \qquad k = 2, \ldots, 19, \tag{45}$$

$$f(0) = 0, \tag{46}$$

$$\left.\frac{df}{d\rho}\right|_{\rho=0} = \bar{v}_{\max}. \tag{47}$$

Equation (45) ensures concavity of f (see, e.g., Mogstad et al. (2018)), Equation (46) ensures that no cars traverse the lane when there are no cars in it, and Equation (47) sets the slope at zero density equal to the group's maximum allowed speed $\bar{v}_{\max}$. Appendix Figure 9 presents an example of an

Appendix Figure F2: Extrapolated flow-density functions



estimated congestion function.

To assign congestion technology to roads outside the two speed-limit classes observed in the estimation sample, we extrapolate the estimated congestion functions to speed limits from 50 to 120 km/h in increments of 10 km/h. For a target speed limit v, we linearly interpolate or extrapolate between the estimated 70 km/h and 90 km/h functions:

$$w = \frac{v - 70}{90 - 70}, \qquad f_v(\rho) = (1 - w) f_{70}(\rho) + w f_{90}(\rho).$$

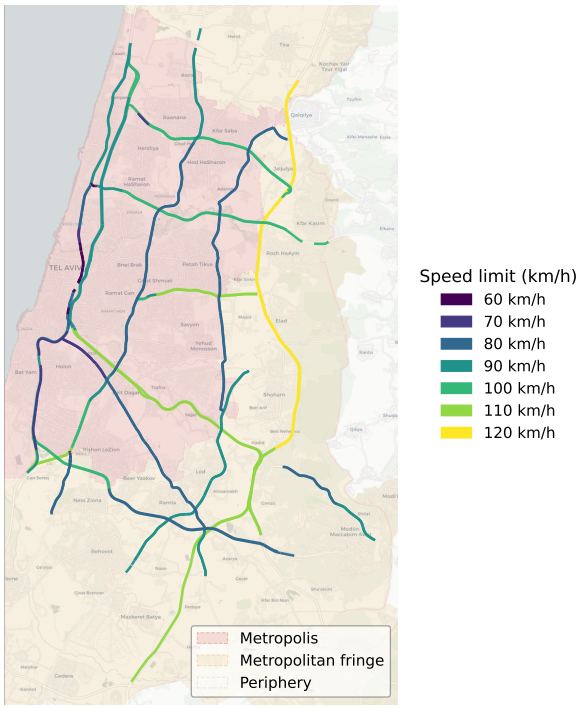
Because this extrapolation is linear, it preserves the shape restrictions imposed on the baseline functions. Appendix Figure F2 visualizes this extrapolation approach. Finally, we convert each lane-level function into a segment-level congestion function using the lane-scaling relationship in Equation (11).

F.2 Road Network Construction

Having estimated congestion technologies for different road classes, we construct the road network used to implement the model in Section 5 for the counterfactual equilibria in Section 7. We begin with the raw OpenStreetMap (OSM) network introduced in Section 3.2 and combine contiguous OSM edges into longer directed segments. The resulting segments are uninterrupted stretches of road between junctions such as intersections and highway entrance or exit ramps.

From this preprocessed network, we identify the major roads in the Central District on which congestion is modeled endogenously. We classify a directed segment as a "highway" edge when it is designated as a motorway or trunk road, has a speed limit of at least 50 kilometers per hour, and lies within the Central District. This procedure identifies 1,047 highway edges. We assign each highway edge the congestion technology estimated in Appendix Section F.1 corresponding to its speed limit, bridge or tunnel status, and number of directional lanes.

Appendix Figure F3: Highway road segment speed limits



We next select the habitual trips introduced in Section 3.3 whose routes traverse at least one highway edge and are represented in the counterfactual demand system, leaving 17,748 trips. For each selected trip, we retain its complete route, including the road segments used to access and leave the Central District highway network. Of the 1,047 highway edges identified above, 1,018 appear on at least one selected route and therefore enter the counterfactual road network, which we visualize in Appendix Figure F3.

For each selected route, we distinguish a prefix, a core, and a suffix. The prefix contains the edges preceding the route's first highway edge, the core extends from the first through the last highway edge, and the suffix contains the remaining edges after the final highway edge. We deduplicate routes with the same sequence of core edges. The 17,748 selected routes, which we refer to as route variants, map to 4,565 distinct core routes. This representation allows route variants with different access or egress portions to share the same congestion-dependent core.

To reduce the number of traversal states further, we also collapse contiguous non-highway edges that lie entirely within the same pricing zone. We collapse edges separately within the prefix, core, and suffix of each route, while leaving highway edges and non-highway edges that cross pricing-zone boundaries separate. For each collapsed segment, we sum the lengths and traversal times of its constituent physical edges.

This procedure yields 9,240 real road segments: 1,018 highway segments assigned the estimated congestion technologies and 8,222 non-highway segments. The non-highway segments continue to contribute to route travel times and remain subject to the applicable location- and time-specific road-use prices. Their speeds are not determined by the equilibrium congestion functions; instead, we construct their transition rates from estimated time-varying free-flow speeds.

To accommodate congestion at core network entry, we add one synthetic overflow/access segment to the beginning of each core route with a "length" of 0.01 kilometers and a free-flow traversal time of one second. Traffic completing a route's prefix enters this segment before moving onto the first physical segment of the core. The overflow segment provides an entry state at which traffic can wait when the downstream core cannot immediately receive all arriving vehicles.

The equilibrium fixed point tracks vehicle masses over the 4,707 edges in the compressed core network.[40] Prefix and suffix edges remain in the calculation of route travel times and prices, but their transition rates do not depend on vehicle density and they therefore do not add states to the equilibrium fixed point.

F.3 Connecting Trip Demand and Traffic Flow Supply

This subsection provides the formal reweighting, traffic-scale calibration, and demand construction summarized in Section 6.2.

Reweighting Trips to Match Cell-Phone-Data-Based Origin-Destination Flows Let i index a habitual trip. Let $o(i)$ and $d(i)$ denote the trip's origin and destination at the coarser level of 250 groups of TAZs at which we observe cell-phone flows (see Section 3.2), and let $g(i)$ denote the driver's home TAZ. We restrict attention to counterfactual-relevant trips with valid origin-destination and home-population information that have at least one trip opportunity during the target hours for the counterfactuals in the relevant weeks. Our objective is to choose a population weight w_i for each supported habitual trip so that the weighted trip panel reproduces the cell-phone-data-based average weekday count N^{Cell}_{odh} of car trips for each origin-destination pair (o, d) and hour h introduced in Section 3.2. Because multiple supported habitual trips can contribute to the same OD-hour count, these external counts generally do not uniquely determine the individual trip weights. We therefore use adult population in drivers' home TAZs to construct a demographic reference allocation and use entropy balancing to select the weights closest to it that still reproduce the external counts.

In particular, let $\mathcal{D}$ denote the distinct calibration dates from February through June of 2021 that we focus our counterfactuals on, and let $R_{i\ell h}$ equal one if trip i is realized on date ℓ in hour h, and zero otherwise. We define the average daily exposure of trip i in hour h as

$$\overline{R}_{ih} \coloneqq \frac{1}{|\mathcal{D}|} \sum_{\ell \in \mathcal{D}} R_{i\ell h}.$$

This daily normalization puts the realized trip exposures on the same average-weekday scale as the external origin-destination counts. If habitual trip i receives population weight w_i, then $w_i \overline{R}_{ih}$ is the average number of population trips in hour h represented by that habitual trip.

For each origin-destination pair (o, d), let $\mathcal{H}_{od}$ denote its supported target hours, let $\mathcal{I}_{od} \coloneqq \{i : o(i) = o, d(i) = d\}$ denote its supported habitual trips, and let $\mathcal{G}$ denote the OD pairs with positive target counts and supported trips. Summing $w_i \overline{R}_{ih}$ over $i \in \mathcal{I}_{od}$ gives the weighted panel's predicted

[40] In the results of the counterfactuals in Section 7, only a small fraction of road segments have binding receiving flow constraints. As such, we deactivate receiving flow constraints for numerical stability.

average weekday count for OD-hour cell (o, d, h). We require this predicted count to equal the corresponding external count N^{Cell}_{odh} for every $h \in \mathcal{H}_{od}$ and $(o, d) \in \mathcal{G}$; we refer to these equality constraints as the OD-hour matching restrictions.[41] The restrictions determine total weighted traffic in each OD-hour cell but generally leave its allocation across individual habitual trips underdetermined.

To allocate traffic across the habitual trips left underdetermined by these restrictions, we choose the weights w_i so that each home polygon's share of weighted drivers is as close as possible to its share of the adult population in a particular distributional sense we introduce below. Let A_g denote the number of residents age 20 or older in home TAZ g, and let m_g denote the number of supported habitual trips whose drivers live in g. We initially assign the adult population of each home TAZ equally across its supported trips by defining

$$b_i := \frac{A_{g(i)}}{m_{g(i)}}.$$

Before imposing the OD-hour matching restrictions, home TAZs contribute in proportion to their adult populations, while supported trips from the same home TAZ are treated symmetrically. We normalize these building blocks b_i over the full supported trip set $\mathcal{I}$ to obtain demographic reference shares

$$p_i := \frac{b_i}{\sum_{j \in \mathcal{I}} b_j}.$$

To convert the reference shares p_i into a benchmark for absolute trip weights, we introduce a common scale $\kappa > 0$, so that κp_i is the benchmark weight for trip i. We then choose the absolute trip weights $\{w_i\}_{i \in \mathcal{I}}$ and κ to minimize their generalized Kullback-Leibler discrepancy from this demographic benchmark subject to reproducing the cell-phone-data-based OD-hour counts:[42]

$$\begin{aligned} \min_{\substack{w_i \geq 0,\ i \in \mathcal{I} \\ \kappa > 0}} \quad & \sum_{i \in \mathcal{I}} \left[w_i \log\left(\frac{w_i}{\kappa p_i}\right) - w_i + \kappa p_i \right] \\ \text{subject to} \quad & \sum_{i \in \mathcal{I}_{od}} w_i \overline{R}_{ih} = N^{\text{Cell}}_{odh}, \qquad h \in \mathcal{H}_{od},\ (o, d) \in \mathcal{G}. \end{aligned} \tag{48}$$

Among the weights that reproduce the external OD-hour flows, the objective selects those that alter the demographic reference allocation as little as possible. The common scale κ allows the total mass of the reference weights to adjust to the absolute scale of the cell-phone counts, while OD-specific tilts

[41] In practice, sparse observed trip-time profiles can make exact hourly matching infeasible. For each OD pair, we first record any positive target hour with no observed habitual-trip support separately. If the remaining hourly targets are still infeasible, we progressively merge adjacent hours and replace their hourly restrictions with an aggregated restriction until the target time distribution lies in the cone generated by the observed exposure profiles. We use the finest grouping of adjacent hours that makes the optimization problem feasible.

[42] This problem formulation has a long history in probability, statistics, and econometrics. The maximum-entropy interpretation of moment-constrained probability choice goes back to Jaynes (1957). With a non-uniform reference distribution, the same logic becomes a minimum-discrimination-information or I-projection problem (Kullback and Leibler, 1951; Kullback, 1997; Csiszár, 1975). The same structure appears in contingency-table raking with fixed margins (Ireland and Kullback, 1968), survey calibration and generalized raking (Deville and Särndal, 1992; Deville et al., 1993), entropy balancing/inverse-probability tilting for estimating causal effects and more general missing data problems (Hainmueller, 2012; Graham et al., 2012). Closely related exponential-tilting estimators use the same information-theoretic logic for econometric moment restrictions (Kitamura and Stutzer, 1997; Imbens et al., 1998).

make the relative adjustments required to satisfy the matching restrictions.

Since Equation (48) is a convex optimization problem with block structure across origin-destination pairs given κ, it can be solved jointly across origin-destination pairs using a block-structured Newton method: each OD pair has its own exponential tilt, but all OD pairs share the scale κ. The resulting w_i is the representative population weight for supported habitual trip i.

Calibrating Traffic Scale and Constructing Empirical Demand The entropy weights place the trip panel on the population scale implied by the cell-phone-data-based origin-destination flows. Because that scale need not coincide with the vehicle-flow scale of the congestion functions estimated in Appendix Section F.1, we use highway sensor volumes to estimate a common factor that maps weighted habitual trips into realized traffic flows. We then combine this factor with the entropy weights and estimated demand primitives to construct the total driver mass Q and type distribution F that enter the equilibrium model in Section 5.

To construct this scaling factor, we first use the network constructed in Appendix Section F.2 to predict sensor-level traffic from the habitual trip panel. Each sensor is mapped to a directed OSM edge. Using the inferred routes constructed in Appendix Section D.1, let M_{is} be an indicator that trip i crosses the edge corresponding to sensor-direction pair s. The trip data record when a driver leaves the origin but not the exact time at which the driver passes a downstream sensor. We therefore aggregate trip departures and observed sensor volumes into 30-minute intervals, the finest resolution at which we can credibly assign a trip's sensor crossing to the same interval as its departure.

Let τ index these intervals, and let $R_{i\ell\tau}$ indicate that trip i is realized on date ℓ with a departure time in interval τ. Under this alignment, the weighted number of habitual trips predicted to pass sensor s in cell (ℓ, τ) is

$$\widehat{C}_{s\ell\tau} = \sum_i M_{is} w_i R_{i\ell\tau}. \tag{49}$$

This prediction uses the full population weight w_i. The weekly availability adjustment introduced below averages counterfactual demand across weeks and therefore does not enter the sensor-flow regression. Let $C_{s\ell\tau}$ denote the total vehicle volume recorded by sensor s over the corresponding 30-minute interval. The calibration sample $\mathcal{C}$ contains sensor-direction-date-interval cells with an observed sensor record, restricted to weeks and within-week intervals in which at least one selected habitual trip has a trip opportunity. We exclude predicted cells without a sensor record since missing sensor observations likely do not correspond to zero traffic.

We estimate a single scale factor, common across sensors and weeks, by regressing realized sensor volumes on predicted habitual-trip crossings without an intercept:

$$\widehat{\zeta} = \frac{\sum_{(s,\ell,\tau)\in\mathcal{C}} \widehat{C}_{s\ell\tau} C_{s\ell\tau}}{\sum_{(s,\ell,\tau)\in\mathcal{C}} \widehat{C}^2_{s\ell\tau}}. \tag{50}$$

The coefficient $\widehat{\zeta}$ maps one weighted predicted crossing into realized vehicle flow.

Having estimated traffic scale, we aggregate potential demand across the calibration weeks. Let $\mathcal{W}$ denote the set of calibration weeks, let O_{iw} be an indicator of whether that trip i has a trip opportunity

in week w, and define its availability share

$$\overline{O}_i \coloneqq \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} O_{iw}.$$

We set the total mass of potential drivers in the model to

$$Q \coloneqq \widehat{\zeta} \sum_{i \in \mathcal{I}} w_i \overline{O}_i. \tag{51}$$

Trips with opportunities in every calibration week contribute their full calibrated mass, while trips available in fewer weeks contribute in proportion to their observed availability. The corresponding population share $\overline{\omega}_i$ of habitual trip i is

$$\overline{\omega}_i \coloneqq \frac{w_i \overline{O}_i}{\sum_{i' \in \mathcal{I}} w_{i'} \overline{O}_{i'}},$$

which supplies the target-population driver weight denoted generically by ω_i in Appendix Section E.4.

To determine the type distribution F, we average the date-varying scheduling preferences estimated in Section 6.1 over the same calibration weeks. Let $\mathcal{L}_{it}(\mathcal{W})$ denote the dates in weeks $\mathcal{W}$ on which departure period t is available for habitual trip i. Writing $\widehat{V}_{i\ell t}(\boldsymbol{\beta})$ for the estimated scheduling preference on date ℓ associated with coefficient vector $\boldsymbol{\beta}$ (which includes trip i's gas cost and value of their best public transit alternative since those are invariant in our model), define

$$\overline{V}_{it}(\boldsymbol{\beta}) \coloneqq \frac{1}{|\mathcal{L}_{it}(\mathcal{W})|} \sum_{\ell \in \mathcal{L}_{it}(\mathcal{W})} \widehat{V}_{i\ell t}(\boldsymbol{\beta}).$$

To represent posterior preference heterogeneity without carrying every draw into the equilibrium calculation, we apply weighted k-means separately to each driver's 500 posterior coefficient draws constructed in Appendix Section E.4, using their posterior importance weights. Let $\widehat{\beta}_{ic}$ and π_{ic} denote the centroid and posterior mass of cluster $c \in \{1, \ldots, K_i\}$, respectively. Combining each centroid with the corresponding habitual-trip attributes and averaged scheduling preferences gives the model type

$$\widehat{\theta}_{ic} \coloneqq \left(\mathcal{J}_i, \mathcal{T}_i, \overline{V}_i(\cdot; \widehat{\beta}_{ic}), \widehat{\alpha}_{ic}, \widehat{\gamma}_{ic} \right).$$

Combining the population trip shares with the within-driver cluster masses gives the discrete type distribution passed to the model:

$$\widehat{F} \coloneqq \sum_{i \in \mathcal{I}} \overline{\omega}_i \sum_{c=1}^{K_i} \pi_{ic} \delta_{\widehat{\theta}_{ic}}, \tag{52}$$

where $\delta_{\widehat{\theta}_{ic}}$ denotes a point mass at type $\widehat{\theta}_{ic}$. Together, Equation (51) and Equation (52) assign mass

$$Q \overline{\omega}_i \pi_{ic} = \widehat{\zeta} w_i \overline{O}_i \pi_{ic}$$

to habitual trip i and representative preference type c. These masses enter the route and departure-

time demand system in Equation (6).

G Equilibrium Computation and Price Optimization

In this appendix, we provide more details about how we discretize the model described in Section 5 to facilitate efficient equilibrium computation and congestion price optimization.

G.1 Model Discretization

First, we explain in detail how we discretize the components of the continuous-time model described in Section 5.

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

$$\bar{q}_{ee't} \coloneqq \Delta^{-1} \int_{t\Delta}^{(t+1)\Delta} q_{ee'}(\tau)\, d\tau, \qquad \bar{q}_{et} \coloneqq \sum_{e' \in \mathcal{E} \cup \{\varnothing\}} \bar{q}_{ee't}, \tag{53}$$

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:

$$\lambda_{ee't} \coloneqq \lambda_{ee'}(\bar{\boldsymbol{q}}_t), \qquad Q_{jt} \coloneqq Q_j(\bar{\boldsymbol{q}}_t). \tag{54}$$

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 in Equation (7) then becomes, on I_t,

$$\frac{d}{d\tau}\boldsymbol{q}_j(\tau) = Q_{jt}\boldsymbol{q}_j(\tau) + a_{jt}\mathbf{e}_1. \tag{55}$$

Forward Mass Propagation Because the generator matrices Q_{jt} are constant within a period, the route-level mass path has a closed-form solution. For local time $u \in [0, \Delta]$,

$$\boldsymbol{q}_j(t\Delta + u) = e^{Q_{jt}u}\boldsymbol{q}_{jt} + a_{jt}\left(\int_0^u e^{Q_{jt}s}\, ds\right)\mathbf{e}_1. \tag{56}$$

It is therefore convenient to define the block operators

$$E_{jt} \coloneqq e^{Q_{jt}\Delta}, \qquad A_{jt} \coloneqq \int_0^\Delta e^{Q_{jt}s}\, ds, \qquad B_{jt} \coloneqq \int_0^\Delta (\Delta - s)e^{Q_{jt}s}\, ds. \tag{57}$$

Evaluating Equation (56) at $u = \Delta$ gives the start-of-next-period recursion,

$$\boldsymbol{q}_{j(t+1)} = E_{jt}\boldsymbol{q}_{jt} + a_{jt}A_{jt}\mathbf{e}_1, \tag{58}$$

while averaging Equation (56) over the period gives the route-level period-average masses

$$\bar{\boldsymbol{q}}_{jt} \coloneqq \Delta^{-1} \int_{I_t} \boldsymbol{q}_j(\tau)\, d\tau = \Delta^{-1} \left[A_{jt} \boldsymbol{q}_{jt} + a_{jt} B_{jt} \mathbf{e}_1 \right]. \tag{59}$$

Backward Value Recursion The duration and price ODEs in Equations (8) and (9) differ in how the relevant quantity accrues but share the same backward structure. Expected duration accumulates continuously while a driver is en route, whereas price is incurred only when the driver completes a segment. We express both objects using a generic continuation value before discretizing them.

Let $X_{jk}(\tau; \boldsymbol{q})$ denote the expected remaining cumulative quantity for a driver occupying the k-th segment of route j at time τ. Suppose that the driver accrues a running amount $r_{jk}(\tau)$ per unit of time while occupying segment k and incurs a lump-sum amount $m_{jk}(\tau)$ upon advancing to segment $k+1$. No quantity remains after trip completion, so $X_{j(K_j+1)}(\tau; \boldsymbol{q}) = 0$. If the driver remains en route at the end of the modeled horizon, we assign terminal continuation value $X_{jk}(T; \boldsymbol{q}) = \bar{X}_{jk}$. Applying the same one-step conditioning used to derive the duration and price equations in the main text gives the generic backward ODE

$$-\frac{d}{d\tau} X_{jk}(\tau; \boldsymbol{q}) = r_{jk}(\tau) + Q_j(\boldsymbol{q}(\tau))[k+1, k]\big(m_{jk}(\tau) + X_{j(k+1)}(\tau; \boldsymbol{q}) - X_{jk}(\tau; \boldsymbol{q})\big), \qquad k \in [K_j]. \tag{60}$$

We now discretize Equation (60). Let r_{jkt} and m_{jkt} denote the running and lump-sum amounts within period t, and define $X_{jkt}(\bar{\boldsymbol{q}}) \coloneqq X_{jk}(t\Delta; \bar{\boldsymbol{q}})$. We collect these start-of-period continuation values across states $k \in \{1, \ldots, K_j + 1\}$ in $\boldsymbol{X}_{jt}(\bar{\boldsymbol{q}})$, with the absorbing component equal to zero, and collect the terminal continuation values in $\bar{\boldsymbol{X}}_j \coloneqq (\bar{X}_{j1}, \ldots, \bar{X}_{jK_j}, 0)'$. The period-$t$ source vector is

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

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 converts Equation (60) to the vector ODE

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

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

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

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

$$\boldsymbol{X}_{jt}(\bar{\boldsymbol{q}}) = A'_{jt} \boldsymbol{c}^X_{jt} + E'_{jt} \boldsymbol{X}_{j(t+1)}(\bar{\boldsymbol{q}}). \tag{61}$$

This recursion is initialized at the horizon by the terminal condition $\boldsymbol{X}_{jM}(\bar{\boldsymbol{q}}) = \bar{\boldsymbol{X}}_j$.

To recover expected duration, we set $r_{jkt} = 1$, $m_{jkt} = 0$, and $\bar{\boldsymbol{X}}_j = \boldsymbol{D}^{\text{ff}}_{jM}$, so

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

To recover expected price, we instead set $r_{jkt} = 0$, $m_{jkt} = p_{e_{jk}t}$, where $p_{et} \coloneqq p_e(t\Delta)$, and $\bar{\boldsymbol{X}}_j = \boldsymbol{0}$. Doing so gives

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

Specializing Equation (61) 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 The continuous-time demand system in Equation (6) assigns each driver a probability density over routes and departure times. To compute demand on the discrete period grid, we replace this continuum of choices with a finite set of route-period alternatives. We assume that each driver's focal time window is a union of periods and let

$$\mathcal{T}_\Delta(\theta) \coloneqq \{t \in \{0, \ldots, M-1\} : I_t \subseteq \mathcal{T}(\theta)\}$$

denote the feasible departure periods for a driver of type θ. Each pair (j,t) with $j \in \mathcal{J}(\theta)$ and $t \in \mathcal{T}_\Delta(\theta)$ represents the choice to take route j and depart at some time during I_t.

Within each period, we approximate scheduling preferences, expected duration, and expected price as constant. Let $\tilde{V}_{jt}(\theta)$ denote the scheduling preference for route j during period t. Using the route-period duration $D_{jt}(\bar{\boldsymbol{q}})$ and price $P_{jt}(\bar{\boldsymbol{q}})$ defined above, the contribution of alternative (j,t) to the numerator of the continuous-time choice probability is therefore

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

The factor Δ converts the continuous-time choice density into the probability mass associated with the entire departure period. Under the unit-scale Type-I extreme-value normalization, we can absorb this factor into utility by defining

$$V_{jt}(\theta) \coloneqq \tilde{V}_{jt}(\theta) + \log \Delta.$$

The resulting route-period utility index is

$$\delta_{jt}(\bar{\boldsymbol{q}}, \theta) \coloneqq V_{jt}(\theta) - \gamma D_{jt}(\bar{\boldsymbol{q}}) - \alpha P_{jt}(\bar{\boldsymbol{q}}), \tag{62}$$

so the probability that a driver of type θ chooses route j and departure period t is

$$s_{jt}(\bar{\boldsymbol{q}}, \theta) \coloneqq \frac{\exp(\delta_{jt}(\bar{\boldsymbol{q}}, \theta))}{1 + \sum_{j' \in \mathcal{J}(\theta)} \sum_{t' \in \mathcal{T}_\Delta(\theta)} \exp(\delta_{j't'}(\bar{\boldsymbol{q}}, \theta))}, \tag{63}$$

for $j \in \mathcal{J}(\theta)$ and $t \in \mathcal{T}_\Delta(\theta)$, with $s_{jt}(\bar{\boldsymbol{q}}, \theta) = 0$ otherwise. Integrating the type-specific choice probability over F gives the population share choosing the alternative:

$$s_{jt}(\bar{\boldsymbol{q}}) \coloneqq \int_\Theta s_{jt}(\bar{\boldsymbol{q}}, \theta)\, dF(\theta). \tag{64}$$

Finally, we assume that drivers choosing (j, t) depart at uniformly random times within I_t. Multiplying their population share by Q and dividing by the period length gives the corresponding within-period inflow rate:

$$a_{jt}(\bar{\boldsymbol{q}}) \coloneqq \frac{Q}{\Delta} s_{jt}(\bar{\boldsymbol{q}}). \tag{65}$$

This inflow rate is the demand object entering the forward mass-propagation equation in Equation (55).

Discretized Equilibrium A discretized equilibrium is a collection $\{\bar{\boldsymbol{q}}_t^*\}_{t=0}^{M-1}$ of period-average movement-level mass vectors defined in Equation (53). Given candidate movement-level masses, the backward recursion in Equation (61) yields expected route-period durations and prices. These objects enter the utility index in Equation (62), which determines type-specific and aggregate route-period shares according to Equations (63) and (64) and route-period inflows according to Equation (65). The resulting inflows propagate route-level masses through the forward recursion in Equation (58), with route-level period averages given by Equation (59). Equilibrium requires that aggregating these route-level period averages across routes reproduce the candidate movement-level average masses:

$$\bar{q}^*_{ee't} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e,\, e_{j(k+1)} = e'\}\, \bar{q}^*_{jkt}, \qquad \forall\, (e, e'),\, t. \tag{66}$$

G.2 Uniformization and Efficient Operator Evaluation

The recursions in Equations (58) and (61) require repeated applications of the operators E_{jt}, A_{jt}, B_{jt} and their adjoints E'_{jt} and A'_{jt} to a small set of vectors. While these operators are not computable in closed form, 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

$$\bar{\lambda}_{jt} \geq \max_{k \in [K_j]} \big(-Q_{jt}[k, k]\big), \qquad \Phi_{jt} \coloneqq I + \bar{\lambda}_{jt}^{-1} Q_{jt}, \qquad \mu_{jt} \coloneqq \bar{\lambda}_{jt} \Delta. \tag{67}$$

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

$$e^{Q_{jt} s} = \sum_{n=0}^{\infty} w_n(\bar{\lambda}_{jt} s) \Phi_{jt}^n, \qquad w_n(\mu) \coloneqq e^{-\mu} \mu^n / n!. \tag{68}$$

Substituting Equation (68) into the integral operators and using the fact that Φ_{jt} does not depend on

s yields

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

and

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

As a result, the operator actions can be written as

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

and the adjoint operators can be written as

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

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

$$F_n = F_{n-1}+w_n, \quad w_{n+1} = w_n\mu_{jt}/(n+1), \quad a_n = \bar{\lambda}_{jt}^{-1}(1-F_n), \quad b_n = \Delta a_n - \frac{n+1}{\bar{\lambda}_{jt}^2}(1-F_n-w_{n+1}). \tag{69}$$

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

$$\mathcal{O}\left(\sum_{t=0}^{M-1} K_j N_{jt}(\varepsilon)\right). \tag{70}$$

G.3 Equilibrium Computation

We solve the discretized fixed point in Equation (66) 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 Equation (54),
$$Q_{jt}^{(r)} \coloneqq Q_j(\bar{\boldsymbol{q}}_t^{(r)}),$$
along with the associated uniformization objects $\bar{\lambda}_{jt}^{(r)}$, $\Phi_{jt}^{(r)}$, and $\mu_{jt}^{(r)}$ from Equation (67).

3. **Backward pass:** starting from $\boldsymbol{D}_{jM}^{(r)} = \boldsymbol{D}_{jM}^{\mathrm{ff}}$ and $\boldsymbol{P}_{jM}^{(r)} = \mathbf{0}$, iterate backward over $t = M-1,\dots,0$ and each route j using the backward recursion in Equation (61) 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 Equations (63)–(65), compute aggregate route-period shares $s_{jt}^{(r)}$ and inflow rates
$$a_{jt}^{(r)} = \frac{Q}{\Delta} s_{jt}^{(r)}.$$

5. **Forward pass:** set $\boldsymbol{q}_{j0}^{(r)} = \mathbf{0}$ for every route j. Then, for $t = 0,\dots,M-1$, use the operators from Equation (57), the forward recursion in Equation (58), and the period-average formula in Equation (59) 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 Equation (66):
$$\bar{q}_{ee't}^{(r,\mathrm{new})} \coloneqq \sum_{j=1}^{J}\sum_{k=1}^{K_j} \mathbb{1}\{e_{jk} = e,\, e_{j(k+1)} = e'\}\, \bar{q}_{jkt}^{(r,\mathrm{new})}.$$

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

and uses the resulting accelerated direction

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

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

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

9. **Convergence test:** stop when

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

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.

G.4 Price Optimization

We now describe how we find a time-varying price schedule that maximizes welfare while accounting for the endogenous traffic equilibrium response.

Continuous-Time Welfare Problem Let $\boldsymbol{\vartheta}$ denote a finite-dimensional vector parameterizing a congestion pricing scheme (e.g. the usage-based and cordon pricing schemes studied in Section 7), and let

$$\boldsymbol{p}(\boldsymbol{\vartheta}) \coloneqq \{p_e(\tau; \boldsymbol{\vartheta})\}_{e \in \mathcal{E},\, \tau \in [0,T]}$$

denote the induced collection of segment-time prices. Let $G(\boldsymbol{q}, \boldsymbol{p})$ denote the equilibrium map from Section 5, now written with its dependence on prices made explicit. For a given parameter vector $\boldsymbol{\vartheta}$, the induced equilibrium movement masses $\boldsymbol{q}^*(\cdot; \boldsymbol{p}(\boldsymbol{\vartheta}))$ satisfy

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

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 $\boldsymbol{q}$ and the candidate price schedule $\boldsymbol{p}$. Let $s_j(\tau; \boldsymbol{q}, \boldsymbol{p}, \theta)$ denote the type-specific route-departure-time choice density from Equation (6) after making the price dependence explicit, and let $s_j(\tau; \boldsymbol{q}, \boldsymbol{p}) \coloneqq \int_\Theta s_j(\tau; \boldsymbol{q}, \boldsymbol{p}, \theta) dF(\theta)$ denote the corresponding aggregate choice density. The aggregate departure rate is then $a_j(\tau; \boldsymbol{q}, \boldsymbol{p}) \coloneqq Q s_j(\tau; \boldsymbol{q}, \boldsymbol{p})$. Using the standard logit inclusive-value formula, population-average consumer surplus and expected toll payments under $\boldsymbol{\vartheta}$ are

$$CS(\boldsymbol{\vartheta}) \coloneqq \int_\Theta \frac{1}{\alpha} \log\left(1 + \sum_{j \in \mathcal{J}(\theta)} \int_{\underline{T}(\theta)}^{\overline{T}(\theta)} \exp(\delta_j(u; \boldsymbol{q}^*(\cdot; \boldsymbol{p}(\boldsymbol{\vartheta})), \boldsymbol{p}(\boldsymbol{\vartheta}), \theta))\, du\right) dF(\theta),$$
$$R(\boldsymbol{\vartheta}) \coloneqq \sum_{j=1}^{J} \int_0^T s_j(\tau; \boldsymbol{q}^*(\cdot; \boldsymbol{p}(\boldsymbol{\vartheta})), \boldsymbol{p}(\boldsymbol{\vartheta})) P_j(\tau; \boldsymbol{q}^*(\cdot; \boldsymbol{p}(\boldsymbol{\vartheta})), \boldsymbol{p}(\boldsymbol{\vartheta}))\, d\tau.$$

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

$$\begin{aligned} \max_{\underline{\boldsymbol{\vartheta}} \le \boldsymbol{\vartheta} \le \overline{\boldsymbol{\vartheta}}} \quad & W_\Lambda(\boldsymbol{\vartheta}) \coloneqq CS(\boldsymbol{\vartheta}) + \Lambda R(\boldsymbol{\vartheta}) \\ \text{s.t.} \quad & G(\boldsymbol{q}^*(\cdot; \boldsymbol{p}(\boldsymbol{\vartheta})), \boldsymbol{p}(\boldsymbol{\vartheta})) = \boldsymbol{q}^*(\cdot; \boldsymbol{p}(\boldsymbol{\vartheta})), \end{aligned} \tag{71}$$

As in Equation (10), $CS(\boldsymbol{\vartheta})$, $R(\boldsymbol{\vartheta})$, and $W_\Lambda(\boldsymbol{\vartheta})$ are averages per potential driver, and $\Lambda \ge 0$ governs the planner's weight on revenue relative to consumer surplus.

Discretized Price Optimization Problem Under the discretization described above, the price path $\boldsymbol{p}(\boldsymbol{\vartheta})$ is represented by the segment-period values $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}) \coloneqq \{p_{et}(\boldsymbol{\vartheta})\}_{e,t}$. In our implementation, $\boldsymbol{\vartheta}$ maps linearly into $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})$; optimizing every segment-period price separately is the special case $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}) = \boldsymbol{\vartheta}$. Given a candidate parameter vector $\boldsymbol{\vartheta}$, the backward recursion for expected prices from Equation (61) 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 Appendix Sections G.1 and G.3. Then the discretized equilibrium associated with $\boldsymbol{\vartheta}$ satisfies

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

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 Equation (64) after making the dependence on candidate prices explicit. The discretized population-average consumer surplus and expected toll payments are therefore

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

The discretized counterpart to Equation (71) is therefore

$$\begin{aligned} \max_{\underline{\boldsymbol{\vartheta}} \leq \boldsymbol{\vartheta} \leq \overline{\boldsymbol{\vartheta}}} \quad & \bar{W}_{\Lambda}(\boldsymbol{\vartheta}) \coloneqq CS(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) + \Lambda R(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) \\ \text{s.t.} \quad & G(\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})), \bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) = \bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})). \end{aligned} \tag{72}$$

Optimization Algorithm We solve the optimization problem in Equation (72) 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 $\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}})$ denote the solution to the discretized equilibrium fixed point. The welfare gradient therefore takes the form

$$\nabla_{\boldsymbol{\vartheta}} \bar{W}_{\Lambda} = \partial_{\boldsymbol{\vartheta}} \bar{W}_{\Lambda} + \partial_{\bar{\boldsymbol{q}}} \bar{W}_{\Lambda} \frac{d\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))}{d\boldsymbol{\vartheta}},$$

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 $d\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))/d\boldsymbol{\vartheta}$, differentiate the fixed-point condition

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

which yields the implicit-function equation

$$[\partial_{\bar{\boldsymbol{q}}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}) - I] \frac{d\bar{\boldsymbol{q}}^*(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))}{d\boldsymbol{\vartheta}} = -\partial_{\boldsymbol{\vartheta}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}(\boldsymbol{\vartheta})), \tag{73}$$

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, doing so would be prohibitively expensive. Instead, we use the adjoint form of the implicit function theorem, which introduces an adjoint vector $\boldsymbol{\lambda}$ satisfying

$$[\partial_{\bar{\boldsymbol{q}}} G(\bar{\boldsymbol{q}}^*, \bar{\boldsymbol{p}}) - I]' \boldsymbol{\lambda} = -\left(\partial_{\bar{\boldsymbol{q}}} \bar{W}_{\Lambda}\right)' \tag{74}$$

and then implies

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

This reformulation is useful 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 like GMRES 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 Section G.3 to obtain $\bar{\boldsymbol{q}}^{*(r)} \coloneqq \bar{\boldsymbol{q}}^{*}(\bar{\boldsymbol{p}}^{(r)})$.

4. **Evaluate welfare:** using $\bar{\boldsymbol{q}}^{*(r)}$, compute the backward price recursion, route-period choice probabilities, and the welfare objective $\bar{W}_{\Lambda}(\boldsymbol{\vartheta}^{(r)})$ from Equation (72).

5. **Evaluate the gradient:** differentiate through the equilibrium map using Equation (73). In practice, we solve the adjoint system Equation (74) 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 $\nabla_{\boldsymbol{\vartheta}} \bar{W}_{\Lambda}(\boldsymbol{\vartheta}^{(r)})$.

6. **Take an L-BFGS-B step:** provide $-\bar{W}_{\Lambda}(\boldsymbol{\vartheta}^{(r)})$ 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 objective becomes sufficiently small; otherwise increment r and return to step 2.

G.5 Λ Threshold for Non-Zero Price Optimality

In addition to computing globally optimal prices for a given revenue weight Λ, we can compute the threshold value of Λ at which the zero-price corner stops satisfying the first-order conditions for a local optimum. The calculation avoids re-solving the full price optimization problem over a grid of Λ values by instead evaluting the one-sided welfare gradient in the feasible parameter directions at the zero-price equilibrium.

Let $\boldsymbol{\vartheta}$ denote the finite-dimensional price parameter vector from Equation (72), normalized so that the lower bound is $\mathbf{0}$.[43] Every feasible local perturbation is represented by a nonnegative direction $\boldsymbol{d} \geq \mathbf{0}$ in the parameter space. Let $\bar{\boldsymbol{q}}^{0} \coloneqq \bar{\boldsymbol{q}}^{*}(\bar{\boldsymbol{p}}(\mathbf{0}))$ denote the discretized zero-price equilibrium.

[43] We apply the threshold calculation only when $\boldsymbol{\vartheta} = \mathbf{0}$ maps into all-zero segment-period prices and each free price cell has a nonnegative scale.

The implemented welfare objective from Equation (72) is

$$\bar{W}_\Lambda(\boldsymbol{\vartheta}) = CS(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})) + \Lambda R(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})),$$

where both components are measured per potential driver. For each parameter m, define the gradient components at zero prices as

$$cs_m \coloneqq \left.\frac{dCS(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))}{d\vartheta_m}\right|_{\boldsymbol{\vartheta}=\mathbf{0}}, \qquad r_m \coloneqq \left.\frac{dR(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta}))}{d\vartheta_m}\right|_{\boldsymbol{\vartheta}=\mathbf{0}}.$$

The one-sided welfare derivative from increasing parameter m is therefore

$$\left.\frac{d\bar{W}_\Lambda(\boldsymbol{\vartheta})}{d\vartheta_m}\right|_{\boldsymbol{\vartheta}=\mathbf{0}} = cs_m + \Lambda r_m. \tag{75}$$

Because the zero-price vector is a lower-bound corner, it satisfies the first-order condition only if the one-sided derivative of welfare is weakly negative in every feasible local direction. We only need to check the coordinate parameter directions because they generate the feasible tangent cone: for any parameter direction $\boldsymbol{d} \geq \mathbf{0}$,

$$\left.\frac{d\bar{W}_\Lambda(s\boldsymbol{d})}{ds}\right|_{s=0^+} = \sum_m d_m(cs_m + \Lambda r_m).$$

Thus, checking all feasible perturbations is equivalent to checking the finitely many coordinate directions: if $cs_m + \Lambda r_m \leq 0$ for every m, every feasible local perturbation has a weakly negative first-order effect; if $cs_m + \Lambda r_m > 0$ for any m, increasing that one parameter raises welfare to first order, so zero prices cannot be locally optimal.

This linearity also gives a parameter-specific threshold. Parameter m becomes welfare improving at

$$\Lambda_m = \begin{cases} 0, & cs_m > 0, \\ \infty, & cs_m \leq 0 \text{ and } r_m \leq 0, \\ -cs_m/r_m, & cs_m \leq 0 \text{ and } r_m > 0. \end{cases}$$

where $\Lambda_m = 0$ means that zero prices already fail the first-order condition when $\Lambda = 0$, and $\Lambda_m = \infty$ means that increasing Λ cannot make parameter m welfare improving because the expected-payment gradient is nonpositive. Since any positive directional derivative is enough to violate the first-order condition at the corner, the reported first-order threshold is therefore

$$\Lambda^* \coloneqq \min_m \Lambda_m.$$

As it happens, computing the two gradient components at zero prices is also simpler than evaluating the general price gradient. Since the expected-price backward recursion has a zero source when $\bar{\boldsymbol{p}} = \mathbf{0}$, $P_{jt}(\bar{\boldsymbol{q}}, \mathbf{0}) = 0$ for every candidate mass vector $\bar{\boldsymbol{q}}$. This implies $\partial_{\bar{\boldsymbol{q}}} P_{jt}(\bar{\boldsymbol{q}}^0, \mathbf{0}) = 0$ and $\partial_{\bar{\boldsymbol{q}}} R(\bar{\boldsymbol{q}}^0, \mathbf{0}) = 0$. The expected-payment gradient therefore does not require differentiating the equilibrium response of traffic: terms involving changes in route shares or equilibrium masses are multiplied by zero expected

prices. Letting $\dot{P}_{jt}^m$ denote the derivative of route-period expected price with respect to ϑ_m, evaluated at fixed masses $\bar{\boldsymbol{q}}^0$ and $\boldsymbol{\vartheta} = \mathbf{0}$, we then have that r_m is the baseline-share average of $\dot{P}_{jt}^m$ across route-period alternatives.

Consumer surplus retains a direct expected-price effect and an indirect congestion effect. Let $s_{jt}^0(\theta)$ denote the zero-price route-period choice probability for type θ, and let $\dot{D}_{jt}^m$ denote the derivative of equilibrium trip duration induced by increasing parameter m. The consumer-surplus gradient is thus given by

$$cs_m = -\int_\Theta \sum_{j,t} s_{jt}^0(\theta)\dot{P}_{jt}^m \, dF(\theta) - \int_\Theta \sum_{j,t} s_{jt}^0(\theta)\frac{\gamma}{\alpha}\dot{D}_{jt}^m \, dF(\theta).$$

The first term is the direct money cost of the marginal price increase. The second term is the equilibrium-duration channel: a marginal price can change route and departure-time choices, which changes movement masses and therefore expected durations. At zero prices, there is no additional consumer-surplus channel through $\partial_{\bar{\boldsymbol{q}}} P_{jt}$ because expected prices are identically zero as a function of $\bar{\boldsymbol{q}}$.

In practice, we compute cs_m using the same implicit-function and adjoint machinery as in the full price optimizer. The zero-price adjoint right-hand side only needs the derivative of consumer surplus with respect to expected durations, because the expected-price-through-equilibrium channel drops out. By contrast, r_m is computed directly from the derivative of expected payments with respect to prices at the zero-price equilibrium.

G.6 Trip Demand Without Network Heterogeneity

Here, we describe how we modify our model to eliminate welfare differences arising from drivers' heterogeneous route and departure-time choice sets while leaving the network loading technology unchanged. The counterfactual has three pieces. First, drivers decide whether to take a trip using the ex-ante value of a randomly drawn inside choice set. Second, conditional on taking a trip, a choice set is drawn and the driver chooses a route and departure time using the same utility index. Third, pricing is restricted to a common access price p paid by any driver who takes a trip, so the planner cannot target particular roads or times directly.

To make the notation explicit, decompose a driver type into

$$\theta = (\eta, \alpha, \gamma), \qquad \eta \coloneqq (\mathcal{J}, \mathcal{T}, V(\cdot)),$$

where η collects the route choice set, focal time window, and scheduling preferences, while (α, γ) are the marginal utility of money and duration coefficient. Let F_η denote the marginal distribution of η and $F_{\alpha,\gamma}$ denote the marginal distribution of (α, γ). Using the marginal distribution F_η rather than the realized η of a driver is what anonymizes choice-set exposure; every coefficient type faces the same lottery over possible inside menus.

For a route j and departure time τ in the feasible set of menu type η, define the anonymized-menu utility index

$$\delta_{Aj}(\tau; \boldsymbol{q}, p, \eta, \alpha, \gamma) \coloneqq V_j(\tau; \eta) - \gamma D_j(\tau; \boldsymbol{q}) - \alpha p,$$

where $D_j(\tau; \boldsymbol{q})$ is still computed by the same backward recursion described in Section 5. For a realized menu η, the inside-choice inclusive value is

$$I(\boldsymbol{q}, p, \eta, \alpha, \gamma) \coloneqq \log\left(\sum_{j \in \mathcal{J}(\eta)} \int_{\underline{T}(\eta)}^{\overline{T}(\eta)} \exp(\delta_{Aj}(u; \boldsymbol{q}, p, \eta, \alpha, \gamma))\, du\right).$$

The ex-ante anonymized inside value for a driver with coefficients (α, γ) is the average of this inclusive value over the menu lottery,

$$\bar{I}(\boldsymbol{q}, p, \alpha, \gamma) \coloneqq \int I(\boldsymbol{q}, p, \eta, \alpha, \gamma)\, dF_\eta(\eta).$$

The probability that a driver with coefficients (α, γ) takes a trip is

$$s_{\text{in}}(\boldsymbol{q}, p, \alpha, \gamma) \coloneqq \frac{\exp\big(\bar{I}(\boldsymbol{q}, p, \alpha, \gamma)\big)}{1 + \exp\big(\bar{I}(\boldsymbol{q}, p, \alpha, \gamma)\big)}.$$

Conditional on taking a trip and drawing menu η, route and departure-time choice is still made using the same route-time utility index:

$$s_{j|\text{in}}(\tau; \boldsymbol{q}, p, \eta, \alpha, \gamma) \coloneqq \frac{\exp(\delta_{Aj}(\tau; \boldsymbol{q}, p, \eta, \alpha, \gamma))}{\sum_{j' \in \mathcal{J}(\eta)} \int_{\underline{T}(\eta)}^{\overline{T}(\eta)} \exp\big(\delta_{Aj'}(u; \boldsymbol{q}, p, \eta, \alpha, \gamma)\big)\, du},$$

for $j \in \mathcal{J}(\eta)$ and $\tau \in [\underline{T}(\eta), \overline{T}(\eta)]$, with $s_{j|\text{in}}(\tau; \boldsymbol{q}, p, \eta, \alpha, \gamma) = 0$ otherwise. Since the access price p is common across all inside alternatives, the price term drops out of this conditional inside-choice probability. Integrating these type-specific choice probabilities gives the aggregate route-departure-time choice density

$$s_{Aj}(\tau; \boldsymbol{q}, p) \coloneqq \int s_{\text{in}}(\boldsymbol{q}, p, \alpha, \gamma) \left[\int s_{j|\text{in}}(\tau; \boldsymbol{q}, p, \eta, \alpha, \gamma)\, dF_\eta(\eta)\right] dF_{\alpha,\gamma}(\alpha, \gamma),$$

where the subscript A denotes the anonymized-menu counterfactual. The route inflow rate is then

$$a_{Aj}(\tau; \boldsymbol{q}, p) \coloneqq Q s_{Aj}(\tau; \boldsymbol{q}, p).$$

Replacing the demand step in the equilibrium map with $a_{Aj}(\tau; \boldsymbol{q}, p)$ defines a modified continuous-time fixed-point map $G_A(\cdot, p)$. The anonymized-menu equilibrium satisfies

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

All other steps of the equilibrium calculation are unchanged: movement masses determine segment traversal rates, the backward recursion computes $D_j(\tau; \boldsymbol{q})$, the demand step computes $a_{Aj}(\tau; \boldsymbol{q}, p)$, the route-level ODE propagates traffic through the network, and route-level masses aggregate back to movement-level masses.

The corresponding money-metric welfare object evaluates drivers before their menu draw. Aver-

aging over potential drivers, consumer surplus and expected access-price payments are therefore

$$CS_A(p) \coloneqq \int \frac{1}{\alpha} \log\big[1 + \exp\big(\bar{I}(\boldsymbol{q}_A^*(\cdot; p), p, \alpha, \gamma)\big)\big] \; dF_{\alpha,\gamma}(\alpha, \gamma),$$
$$R_A(p) \coloneqq p \int s_{\text{in}}(\boldsymbol{q}_A^*(\cdot; p), p, \alpha, \gamma) \, dF_{\alpha,\gamma}(\alpha, \gamma).$$

In $CS_A(p)$, $\bar{I}$ contains the common access price and the option value of the route and departure-time choice within a randomly drawn inside menu, while the outer log-sum captures the option value of choosing whether to take a trip at all. The population-average welfare objective with revenue weight Λ is then

$$W_{A,\Lambda}(p) \coloneqq CS_A(p) + \Lambda R_A(p),$$

where $CS_A(p)$, $R_A(p)$, and $W_{A,\Lambda}(p)$ are averages per potential driver, as in the general price optimization problem. Pricing counterfactuals compare $W_{A,\Lambda}(p)$ to $W_{A,\Lambda}(0)$, and the restricted optimal common price solves $\max_{p \in [\underline{p}, \bar{p}]} W_{A,\Lambda}(p)$.

This construction preserves the traffic supply side and conditional route assignment induced by expected durations but removes welfare differences caused by a driver's realized choice-set exposure. Heterogeneity in the welfare object remains only through (α, γ) and how those coefficients value the common ex-ante menu lottery and common access price. Because the price is the same for any trip, the planner cannot use the price instrument to target particular roads, departure periods, or routes.