Value Pricing or Lexus Lanes?
The Distributional Effects of Dynamic Tolling[*]

Cody Cook[†]   Pearl Z. Li[‡]

Job Market Paper

December 9, 2023

Click here for most recent version

Abstract

This paper studies the welfare and distributional effects of dynamically priced highway toll lanes. To quantify the equilibrium effects of tolling, we develop and estimate a model of driver demand, the road technology, and the pricing algorithm. The demand model features heterogeneous drivers choosing both where and when to drive, as well as uncertainty about prices and travel times. A key welfare channel is the option value of tolling: even drivers who infrequently take the priced lanes can benefit from having the option but not the obligation to pay for speed. The model is estimated using data on toll transactions, historical traffic conditions, and driver characteristics from the I-405 Express Toll Lanes in Washington State. Relative to a world in which the same number of highway lanes are all free, status-quo tolling increases aggregate welfare and benefits drivers in all income quartiles, driven in large part by the option value. Moreover, we find that drivers in the bottom income quartile gain the most under status-quo tolling. Low-income drivers have the longest I-405 commutes and they face low prices relative to their time savings from the priced lanes. They also have high option values of tolling because they are more price-sensitive, so they are more likely to be marginal when deciding between the priced and unpriced lanes. Finally, we show how simple revisions to the pricing algorithm can increase aggregate welfare and achieve redistributive goals.

1 Introduction

Traffic congestion has large human and environmental costs, resulting in substantial wasted time, wasted fuel, and air pollution. As early as Pigou (1920), economists have advocated congestion pricing as a way of correcting this negative externality. In the past twenty years, cities around the world have begun charging congestion fees for driving in crowded downtown zones.[1] In contrast, today in the United States, the predominant form of congestion pricing takes place on highways, where dynamically priced toll lanes run adjacent to existing unpriced lanes.[2] By updating prices in response to real-time traffic conditions, these toll lanes keep speeds high and offer paying drivers substantial time savings. Drivers have the option but not the obligation to pay for speed, as they retain the unpriced lanes as an alternative.

However, there is disagreement about the distributional effects of highway toll lanes. On one side, policymakers refer to dynamic tolling as "value pricing" and emphasize that it provides choice to drivers (Samdahl et al., 2013). On the other side, opponents are concerned that "Lexus lanes" enrich the wealthy at everyone else's expense (Astor, 2017; Rosendorf, 2018). Evaluation of these perspectives depends on two empirical objects: the distribution of driver preferences and what we call the "road technology"—the relationship between traffic quantities and travel times. When one lane becomes tolled, drivers substitute from the newly priced lane into the remaining unpriced ones, increasing travel times in the unpriced lanes. High peak-hour prices may also induce drivers to substitute toward driving off-peak (or not at all), which can increase average speeds when the road technology is convex. Finally, since tolling changes the *predictability* of travel times, having the option to take the priced lanes can serve as insurance against worse-than-expected traffic conditions.

In this paper, we study the aggregate and distributional impacts of dynamic tolling. To do this, we bring together data on toll transactions, historical traffic conditions, and driver characteristics from the I-405 Express Toll Lanes in Washington State. We begin by presenting two sets of descriptive facts: first, aggregate speed and throughput increased after the introduction of tolling on this highway, and second, low-income drivers face advantageous trade-offs between price and travel time savings in the toll lanes. Next, to quantify the equilibrium effects of tolling, we build and estimate a model of driver demand, the road technology, and the pricing algorithm. In particular, the demand model incorporates the features of dynamic tolling highlighted above: choices of where and when to drive, as well as uncertainty about prices and travel times. Using the estimated model, we find that low-income drivers in fact gain the most from status-quo tolling, and we explore how equilibrium outcomes would change under counterfactual pricing policies.

Our empirical analysis uses data from Interstate 405 in Washington State. The tolled part of this corridor features both unpriced general-purpose (GP) lanes and priced high-occupancy/toll (HOT) lanes. Prices are updated dynamically every five minutes and vary across "trip definitions"—where

---

[1]Cities that have adopted downtown congestion pricing include Singapore (Phang and Toh, 2004), London (Leape, 2006), and Stockholm (Börjesson et al., 2012). New York City plans to do so within a few years (Ley, 2023).

[2]Wood et al. (2021) report that at the end of 2019, there were 53 highways with toll lanes in the United States, totaling 1858 lane-miles across 16 states. (These numbers do not include toll roads, on which all lanes are priced.) Of the 53 toll lane facilities, 40 used dynamic pricing.

the driver enters and exits the toll lanes—motivating our definition of a market as a group of highway entry-exit pairs with the same trip definition. Our primary dataset is the universe of I-405 toll transactions in 2019, which we link to proxies for driver income. We additionally construct a panel of historical traffic conditions, which vary at the market and five-minute interval level.

We begin by investigating long-run changes in GP and HOT speed and throughput around September 2015, when the previously carpool-only lanes were converted to HOT lanes. We document that aggregate speed and throughput increased in both the always-unpriced GP lanes and the newly priced HOT lanes in the years after the HOT opening. In particular, average throughput grew by a modest 5 percent in the GP lanes and a more substantial one-third in the HOT lanes. However, multiple policy changes were bundled together in the first year after the HOT opening, including the construction of an additional HOT lane, making it difficult to isolate the effect of pricing. These limitations in part motivate the need for our structural model.

Next, we show suggestive evidence that it is *low-income* drivers who gain the most from tolling. We use two proxies for driver income, the median household income in her home Census tract and the retail price of her car. By both measures, lower-income drivers tend to belong to "longer" I-405 markets—that is, their travel involves greater distances on I-405. Drivers in these longer markets naturally face higher-mean and higher-variance travel times in both the GP and HOT lanes, as well as higher HOT time savings, but they do not face proportionally higher prices. Moreover, patterns in the toll transaction data suggest that higher-income drivers are willing to pay more to avoid travel time, but that preference heterogeneity along observables is modest. Among peak-hour HOT trips taken in the same market and hour of day, those taken by higher-income drivers have very slightly lower time savings and very slightly higher prices.

Since a full welfare analysis requires accounting for the equilibrium effects of dynamic tolling, our next step is to develop a model of driver demand, the road technology, and the pricing algorithm. The model provides a framework for decomposing gains and losses, both across drivers and across channels; it also allows us to simulate welfare under counterfactual policies. Our structural model restricts attention to the morning commute, a setting where drivers are likely to make repeated choices with stable preferences.

On the demand side, we extend the classical Vickrey (1969) trip scheduling model to incorporate a new welfare channel, the option value of tolling. Drivers in the model have heterogeneous preferences over prices, travel times, and how early or late they are to their destinations, relative to their ideal arrival times. They make decisions in two stages. In the first stage, each driver chooses a departure time (or the non-405 outside option) given imperfect information about future prices and travel times. In the second stage, she observes the realized price and travel times at her chosen departure time and decides between the priced and unpriced routes. Even drivers who rarely choose the priced route may still derive value from having the option to do so—for example, by choosing "riskier" departure times, knowing that they can pay for speed in the event of worse-than-expected traffic.

To close the model, in equilibrium, travel times are determined by the corridor's road technology

and prices are set by a pricing algorithm. The road technology model embeds a static relationship between traffic density and traffic speed inside a discretized model of traffic dynamics in space and time. Dynamics are important because drivers from multiple markets (groups of highway entry-exit pairs) and multiple departure times contribute to the traffic density, and hence the traffic speed, of a given road segment during a given time interval. Travel times in a given market and at a given departure time depend on speeds along the driver's entire trajectory in spacetime. We model the pricing algorithm as a flexible function from GP and HOT travel times into prices for each market. This is an approximation of the true pricing algorithm, which we observe but which is difficult to use directly in counterfactual simulations.[3]

We estimate the two stages of the demand model jointly using the simulated method of moments. Our estimation approach is designed to overcome a key data limitation: we don't observe market shares of either departure times or routes at the market level. The mean preference coefficients on price and travel time are identified using variation from price rounding and car crashes, which shift prices and travel times when the driver is already on the road. We also use morning precipitation, which increases the variance of prices and travel times, as a beliefs shifter to identify drivers' costs of being early and late to their destinations. Intuitively, if drivers' disutility of being late is high, they will "buy more insurance" by shifting their departure times earlier on mornings when it rains or snows. Finally, we identify the parameters governing heterogeneous preferences for price and travel time by matching micro moments in the toll transaction data.

We estimate that drivers have moderate preference heterogeneity and low scheduling costs. The average driver is willing to pay $20.40 to avoid one hour of travel time, holding fixed her time early and late. This value of travel time (VOTT) increases in the driver's tract income and slightly decreases in her car price. The 5th and 95th percentiles of the VOTT are $17.16 and $22.59, respectively.[4] Scheduling costs are relatively low, with the average driver willing to pay $3.53 to avoid being one hour early and $4.33 to avoid being one hour late relative to her ideal arrival time.

We find that status-quo tolling is welfare-improving in aggregate, relative to a no-toll counterfactual in which there is the same total number of lanes, but all lanes are free. While total driver surplus (during the morning commute) decreases by $29,100 per day, daily toll revenue increases from zero to $37,600 per day, so that net welfare increases by $8,500 per day, assuming full recycling of toll revenue. Under status-quo tolling, conditional on taking I-405 (versus choosing the non-405 outside option), drivers pay an average of 77 cents per day in tolls, but are partially compensated by an average of 1.4 fewer minutes of travel time, 2.2 fewer minutes early, and 0.9 fewer minutes late to their destinations. These time gains arise because tolling increases substitution to the outside option by about 3 percentage points. This substitution away from I-405 is greatest during peak

---

[3]However, we do use the true algorithm and algorithm inputs to reconstruct the unrounded prices. These unrounded prices are then used to estimate both the price rounding regression discontinuity and the pricing algorithm approximation.

[4]A common benchmark for the value of travel time comes from the U.S. Department of Transportation, which estimates the value of time spent on local personal travel (including commuting) as 50 percent of the median hourly wage (Belenky, 2011). Our mean estimate is 62 percent of the $32.91 median hourly wage in the Seattle-Tacoma-Bellevue metropolitan statistical area in May 2019 (Bureau of Labor Statistics, 2020).

hours from 6–10 AM; we find little substitution toward driving off-peak, during the very early or very late morning.

Moreover, status-quo tolling increases welfare in all income quartiles, and the greatest gains (+$5,000 per day) accrue to drivers in the bottom income quartile. These considerable bottom-quartile gains are driven in large part by geography. We find that relatively high-income drivers in short markets are net losers under status-quo tolling, while relatively low-income drivers in long markets are net winners. The magnitudes of surplus being transferred are substantial: in aggregate, drivers in one of the shortest markets are made worse off by about $3,200 per day, while drivers in the longest market are made better off by about $8,500 per day—almost identical to the aggregate welfare gain obtained from summing across all markets.

Next, we decompose the welfare changes from tolling into two channels—ex ante value and option value—and find that the option value is an important driver of both aggregate and low-income welfare gains. To do this, we compute welfare in an intermediate step, supposing that drivers face status-quo prices but must make both their departure time and route choices in the first stage, before they observe the realizations of prices and travel times. The ex ante value of tolling accounts for the re-equilibration of prices and travel times and could therefore, in principle, be positive or negative; we estimate its aggregate amount to be about $1,600 per day. The option value, which is positive by construction, contributes an additional $6,900 per day. Low-income drivers benefit from both high ex ante values (due to geography) and high option values (due to their lower willingness to pay to avoid travel time). While the option value takes roughly similar dollar amounts across markets, the total amount accruing to drivers in the bottom half of the income distribution is almost double that of drivers in the top half.

Finally, we simulate equilibrium outcomes under alternative pricing policies. Raising the price ceiling from $10 to $12 is welfare-improving both in aggregate (increasing aggregate welfare by an additional $1,700 per day relative to the status quo) and across income quartiles.[5] The higher price ceiling allows the algorithm more flexibility with which to manage high-traffic periods. While lower-income drivers in longer markets face slightly higher prices than under the status quo, these prices are still far from being proportional to the HOT time savings. We also explore two forms of income-based pricing that are under consideration by WSDOT: a 50 percent proportional discount and a $2 flat discount for low-income drivers (Washington State Transportation Commission, 2021). We find that both forms of income-based pricing reduce toll revenues and increase HOT travel times, reducing aggregate welfare relative to the status quo. However, to the extent that transportation costs (in both time and money) already fall most heavily on low-income drivers, policymakers may still value the redistributive potential of income-based pricing.

Related literature   This paper contributes to several strands of the literature.

First, our paper contributes novel empirical evidence on the efficiency and equity of highway toll lanes. A rich theoretical literature has explored the heterogeneous welfare effects of tolling,

---

[5]The existing algorithm generates unrounded prices between 50 cents and $12. We consider a $12 price ceiling in order to avoid extrapolating our approximation of the pricing algorithm beyond the data.

extending the Vickrey (1969) model of endogenous trip scheduling to incorporate heterogeneous drivers (Arnott, de Palma and Lindsey, 1994), "value pricing" of a subset of the highway (Small and Yan, 2001; Verhoef and Small, 2004), and alternate forms of congestion (Hall, 2018). Recent empirical work has used structural methods to analyze the distributional effects of tolling using data from household travel surveys (Hall, 2020) and toll transactions (Mattia, 2022).[6] Our paper extends this literature in two ways. First, we link the rich transaction-level data on driver decisions to characteristics of the same drivers. This allows us to explore heterogeneous welfare effects along several dimensions, including (observable proxies for) income, geography, and the value of travel time. Second, we model an additional welfare channel, the option value of tolling, which we estimate to be quantitatively important, both in aggregate and for low-income drivers.

A broader empirical literature studies the effectiveness and incidence of congestion pricing in urban settings, rather than on highways. Research in this area has evaluated real-world urban congestion pricing (Phang and Toh, 2004; Leape, 2006; Börjesson et al., 2012) and quantified the aggregate welfare effects of counterfactual cordon pricing (Tarduno, 2022), time-of-day pricing (Kreindler, 2023), and congestion surcharges on taxis and ridesharing (Arora, Zheng and Girotra, 2020; Rosaia, 2020). Turning to the distributional effects of urban congestion pricing, two recent papers have evaluated cordon pricing (Herzog, 2020) and distance-based pricing (Barwick et al., 2021) in longer-run equilibria with residential and mode (e.g., driving vs. public transit) choices. They find that low-income commuters gain the most from these other forms of congestion pricing, mirroring the results in our paper.[7]

Finally, another closely related area of the literature uses variation from highway toll lanes to estimate driver preferences for travel time and reliability. Small (2012) reviews empirical estimates of the value of travel time, an important input for valuing the time savings from both road pricing and infrastructure improvements.[8] By directly modeling drivers' scheduling costs, we estimate the distribution of the value of travel time holding fixed how early or late each driver is to her destination. Moreover, it is understood that drivers care not just about mean travel times, but also about their variance. Papers which estimate this value of reliability (Lam and Small, 2001; Small, Winston and Yan, 2005; Brent and Gross, 2018) and the related value of urgency (Bento, Roth and Waxman, 2020) have historically treated reliability as a reduced-form product characteristic entering drivers' route choice decisions. Our two-stage demand model generates this preference for reliability from drivers' first-stage departure time choice under imperfect information: drivers value reliability because it reduces their realized scheduling costs in the second, route choice, stage.

Overview The remainder of the paper is organized as follows. Section 2 provides background on how dynamic tolling works in our empirical setting. Section 3 describes our primary data sources. Section 4 presents descriptive evidence on the aggregate and distributional effects of tolling. The

---

[6]Hallenbeck et al. (2019) also conduct descriptive analysis of distributional effects using toll transaction data.

[7]Other papers have estimated that a distance-based tax (i.e., a tax on vehicle miles traveled) would be regressive, but less regressive than existing gas taxes (Martin and Thornton, 2018; Glaeser, Gorback and Poterba, 2022).

[8]Recent papers have also estimated the value of travel time in other settings, including residential choice (Su, 2019) and ridesharing (Buchholz et al., 2020; Goldszmidt et al., 2020).

equilibrium model of driver demand, the road technology, and the pricing algorithm is presented in Section 5 and estimated in Section 6. In Section 7, we evaluate welfare and other outcomes under status-quo tolling and counterfactual pricing policies. Section 8 concludes.

2   The I-405 Express Toll Lanes

Our empirical setting is Interstate 405 (I-405) in Washington State, a well-trafficked highway which connects several Seattle-area suburbs. Its tolled section, mapped in Figure B.15, runs north-south for seventeen miles between Lynnwood and Bellevue, east of and across Lake Washington from Seattle proper. Along this corridor, there are typically one to two high-occupancy/toll (HOT) lanes and three to four unpriced general-purpose (GP) lanes in each direction. The priced lanes run directly adjacent to the unpriced lanes, separated by double white lines.

The highway is designed so that drivers have a relatively accurate picture of current prices and travel times (Figure A.1). Drivers may only enter or exit the toll lanes at designated access points, which are typically available between every few interchanges, roughly every few miles.[9] As drivers approach each access point, they are shown the current prices on two sequential electronic signs. Some but not all HOT access points also have electronic signs which show estimated travel times in both the priced and unpriced lanes.

Prices vary dynamically across time, and are subject to both rounding and a floor and ceiling. Tolling is in effect from 5 AM to 7 PM on weekdays, excluding major federal holidays.[10] The pricing algorithm is designed to maintain HOT speeds of at least 45 miles per hour for 90 percent of tolled hours. That is, the objective of I-405 tolling is explicitly to reduce congestion rather than to raise revenue. New prices are computed every five minutes using data on speed and throughput from induction loops embedded in the pavement.[11] Importantly for our identification strategy, while the pricing algorithm initially calculates continuous toll rates, the prices actually faced by drivers are rounded to the nearest 25 cents. Prices are also subject to both a 75-cent floor and a $10 ceiling; we explore raising the ceiling in one of our policy counterfactuals.

Within a five-minute interval, prices also vary across "trip definitions"—where the driver enters and exits the toll lanes—which form the basis of our market definition. The algorithm independently computes prices for each trip definition. In particular, prices are non-additive across road segments, so that the A to C toll is not necessarily equal to the sum of the A to B toll and the B to C toll. We therefore define a market as a group of highway entry-exit pairs which share the same trip definition.[12] In each direction, there are fourteen markets, of which three do not have a feasible

---

[9] In practice, drivers sometimes illegally cross the double white lines separating the GP and HOT lanes rather than entering and exiting solely at the designated access points. Most of the access points are short segments on which the double lines are replaced with single dashed lines. There are also two ramps which provide direct access to the HOT lanes from an interchange.

[10] Outside these hours, even single-occupancy vehicles can drive in the HOT lanes for free.

[11] Dynamic pricing is also used in other settings, including Uber surge pricing (Castillo, 2019; Castillo, Knoepfle and Weyl, 2023) and revenue management in the airline (Williams, 2021), railroad (D'Haultfœuille et al., 2022), and hotel (Cho et al., 2018) industries.

[12] We assume that "the" trip definition for a given market is the one in which drivers take the HOT lanes for

HOT route. Appendix B.1 provides more details.

Paying vehicles have two ways to pay tolls, though we abstract away from this distinction in the model. Most drivers pay using online accounts which are linked to transponders inside their vehicles. However, drivers without transponders can still take the HOT lanes: their license plates are photographed and they receive bills in the mail. An additional $2 fee is charged for "paying by plate" in this way. Non-paying vehicles include buses, motorcycles, and high-occupancy vehicles (HOV), which can drive in the HOT lanes for free. Drivers in this last group must have their transponders set to the HOV setting.[13]

While our primary interest is in comparing the status quo to a counterfactual where all lanes are general-purpose, in the case of I-405, HOT lanes actually replaced HOV rather than GP lanes. In fact, the opening of the I-405 HOT lanes in September 2015 bundled together several policy changes. First, along the entire corridor from Lynnwood to Bellevue, previously HOV-only lanes were converted to HOV-plus-toll (i.e., HOT) lanes. Second, an entirely new HOT lane was built between Woodinville and Bellevue, on the higher-trafficked southern half of the newly tolled section. Third, occupancy requirements were increased. Vehicles with two occupants, which were previously permitted to access the HOV lanes for free, were now required to pay the same toll as single-occupancy vehicles during weekday peak hours. Vehicles with three or more occupants continued to drive for free at all times.[14]

3 Data

We combine data on toll transactions, historical traffic conditions, and driver characteristics. Several datasets were obtained in partnership with the Washington State Department of Transportation (WSDOT) Tolling Division, which administers the I-405 HOT lanes. Our primary estimation sample is the calendar year 2019, the intersection of our data samples. Appendix B provides additional details on data construction.

3.1 Toll transactions

We obtain data on the universe of I-405 HOT transactions in 2019 from WSDOT. For each transaction, we observe the trip definition, HOV status, price paid, and timestamps for each toll gantry driven under during the trip. Each transaction is linked to an account identifier; we match a subset of accounts to the Census tract of the account holder's billing address and the make, model, and year of her vehicle.[15]

---

as long as possible between their highway entry and exit. Each trip definition is associated with a unique market. Most markets are associated with a single trip definition, but there are two exceptions in which there is a nontrivial trade-off between two trip definitions, where one typically has both higher prices and higher HOT time savings. We say that in these two markets, there are two HOT routes to choose from.

[13]In 2019, 41 percent of paid HOT transactions are paid by plate and 31 percent of all transactions are HOV.

[14]Initially, the HOT lanes were tolled at all hours, including nights, weekends, and holidays. The current tolled hours were introduced in March 2016.

[15]52 percent of HOT drivers are matched to a tract income; 39 percent are matched to a car price.

Table 1: Paid transaction summary statistics

	All			Peak hours			Peak hours + income proxies		
	p25	p50	p75	p25	p50	p75	p25	p50	p75
Trip attributes									
Time saved (mins)	0.153	1.928	6.903	0.688	4.217	9.229	0.779	4.398	9.465
Price paid ($)	0.750	0.750	4.750	0.750	2.750	6.500	0.750	3.000	6.500
Price per min saved	0.467	0.914	5.694	0.459	0.733	2.107	0.457	0.721	1.949
Counts									
Transactions		9,633,458			6,160,687			4,940,253	
Unique drivers		1,123,028			746,706			530,000	

Note: Each observation is a paid toll transaction. Southbound peak hours are 5 AM to 11 AM; northbound peak hours are from 1 PM to 7 PM. The third set of columns, labeled "Peak hours + income proxies," restricts to paid, peak-hour transactions which are matched to at least one of a tract income or a car price.

Table 1 reports summary statistics of paid transactions. Our analysis restricts attention to peak hours, which we define as southbound 5 AM to 11 AM and northbound 1 PM to 7 PM. In this sample of 6.1 million transactions by about 747,000 unique drivers, the median time saved in the HOT lanes is 4.2 minutes and the median price paid is $2.75. We identify the preference heterogeneity parameters in the structural model by matching empirical covariances between income proxies and these trip attributes. Further restricting to transactions that are matched to a tract income or a car price results in 4.9 million transactions by 530,000 unique drivers.

3.2 Historical traffic conditions

We construct a panel of historical traffic conditions at the five-minute level and a panel of potential market sizes at the daily level. Together, these two datasets describe the inputs to driver decision-making.

Traffic conditions vary substantially across markets (groups of highway entry-exit pairs). Table A.1 reports market lengths and median travel times and prices in each southbound market, separately for peak and off-peak hours. Figure A.2 shows how travel times and prices vary by time of day and across days in two example markets, one short and one long.

Prices and pricing algorithm inputs We obtain historical rate cards and pricing algorithm inputs from September 2015 to March 2022 from WSDOT. The rate cards show realized prices which have been rounded to the nearest 25 cents. Since WSDOT does not record historical pre-rounding, algorithm-generated prices, we reconstruct them from the algorithm source code, also shared by WSDOT. The algorithm inputs are a processed version of the induction loop data, which we describe next.

Traffic sensor data We add I-405 traffic sensor data from January 2011 to December 2021, downloaded from the Washington State Transportation Center (TRAC)'s online TRACFLOW tool. The sensors are induction loops embedded in the each lane, spaced out roughly every half mile on the mainline, as well as on highway on- and off-ramps.

Each loop reports its average speed and throughput in each five-minute interval.[16] The loop's throughput is the number of vehicles passing over it per unit of time. We are also interested in traffic density, the number of vehicles per unit of distance, which the loops do not measure directly. We approximate density (cars per lane per mile) at each loop in each five-minute interval by dividing throughput (cars per lane per hour) by speed (miles per hour), following the transportation literature (Hall, 2005). Appendix B.2 validates this approximation empirically; Appendix C.2 discusses the theoretical relationships between speed, density, and throughput.

To turn speeds on road segments into market- and route-level travel times, we assume that travel speed is constant within each segment (between two adjacent loops) and five-minute interval. When computing HOT travel times, we additionally account for time spent in the GP lanes, for example between the highway on-ramp and the HOT access point. Since markets group multiple highway entry-exit pairs, we define "the" GP or HOT travel time in a market as the travel time between the first highway entry and the last highway exit in the market.

Lastly, to compute the quantity of drivers departing (i.e., entering I-405) in each five-minute interval in each *market origin* (not market), we sum the throughputs on all on-ramps for that market origin.[17]

Tract-to-tract travel flows To construct potential market sizes, we use historical travel flows between ordered pairs of Census tracts aggregated from cell phone GPS pings in 2019. These flows include all forms of travel between a given tract pair—not only travel on I-405 or even only travel by car. We use GraphHopper's routing API to group together all trips for which a particular entry-exit pair on I-405 was one of the top three suggested routes. This procedure collapses *realized* flows between Census tracts into *potential* flows between I-405 entries and exits, which become our estimates of potential market sizes. Appendix B.4 discusses in more detail; Figure B.18 illustrates the procedure for the southbound full-length market. The resulting market sizes are at the (market, date) level.

Other price and travel time shifters We use car crashes as additional shifters of *realized* prices and travel times and precipitation as a shifter of drivers' *beliefs* about prices and travel times. The data on car crashes, obtained via a public disclosure request from WSDOT, contain

---

[16]The loop speeds are top-coded at 60 miles per hour; we replace top-coded speeds with freeflow speeds estimated from the toll transaction data, which vary at the (loop, day of week, hour of day) level.

[17]An added complication is that each on-ramp contains both metered (general-purpose) and unmetered (carpool-only) lanes, and the metered throughputs reflect rationed demand for departure times. Within each day and ramp, we reallocate the metered throughputs so that they match the profile of unmetered throughputs. That is, we assume that single-occupancy drivers demand the same distribution of departure times as carpooling drivers. Appendix B.5 discusses this procedure and assumption in more detail.

the date and time, the milepost, the location of first impact (i.e., which lane or shoulder, which we use to determine whether the crash happened in the GP or HOT lanes), and other characteristics for each crash in 2019. The precipitation data were collected by Automated Surface Observing Systems (ASOS) and downloaded from Iowa State University's Environmental Mesonet. We use hourly precipitation (in inches) at three Seattle-area weather stations in 2019.[18]

3.3 Driver characteristics

Our primary driver characteristics are Census tract median household incomes and car prices, two proxies for income. We take tract-level demographics, including income distributions, from the 2019 American Community Surveys (ACS). We estimate the manufacturer-suggested retail price (MSRP) for each combination of car make, model, and year using data on vehicle registrations from March 2017 to December 2022 from the Washington State Department of Licensing (WSDOL). Appendix B.3 describes our procedure for estimating vehicle MSRPs.

We construct two samples of drivers, the HOT sample and the unconditional sample. The HOT sample contains real-world drivers who are observed taking the I-405 toll lanes at least once in 2019. These real-world drivers are linked to their tract incomes and car prices via their billing addresses and car make-model-years. The unconditional sample contains simulated drivers, drawn from the joint distribution of tract incomes and car prices in the population of potential I-405 drivers. To construct this joint distribution, we sample Census tracts, weighting by tract-to-tract travel flows, then sample vehicle registrations conditional on tract; Appendix B.4 describes this procedure in more detail. Figure 1 shows that the tract incomes and car prices of drivers taking the HOT lanes are very similar to those in the broader 405-eligible population.

4 Descriptive evidence

In this section, we present two pieces of descriptive evidence on the aggregate and heterogeneous effects of dynamic tolling. First, we document increases in aggregate speed and throughput in both the GP and HOV/HOT lanes after the opening of HOT lanes, though we caution that these gains can't be attributed entirely to the introduction of pricing. Next, we show suggestive evidence that low-income drivers may be the greatest beneficiaries from tolling in our setting. These descriptive facts motivate the addition of the structural model and will provide intuition for the results of our welfare analysis.

4.1 Long-run speed and throughput

We begin by exploring how speed and throughput have changed in the long run, both before and after the HOT lanes opened in September 2015. The goal is to document what actually happened

---

[18]ASOS is jointly operated by the National Weather Service, the Federal Aviation Administration, and the Department of Defense. The three weather stations are in Everett (about 11 miles north of I-405's northern terminus in Lynnwood), Renton (about 8.5 miles south of where I-405 tolling ends in Bellevue), and Seattle-Tacoma International Airport (about 13 miles southwest of Bellevue).

Figure 1: Distributions of income proxies in all drivers vs. HOT drivers

(a) Tract income   (b) Car price
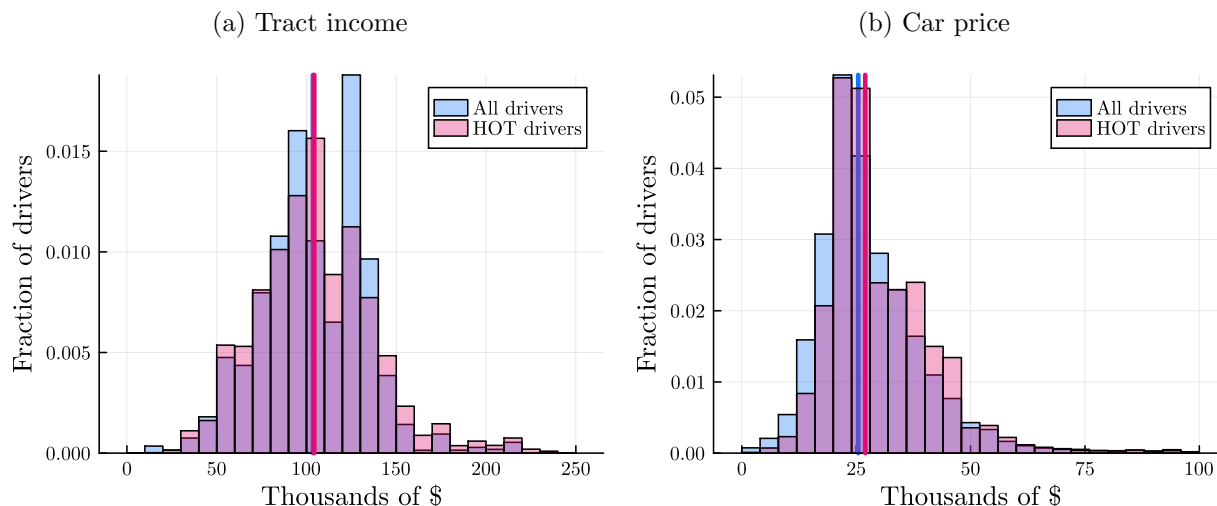
Note: Histograms show marginal distributions of income proxies in the unconditional sample of potential I-405 drivers (blue) vs. the HOT sample of drivers who are observed taking the I-405 toll lanes at least once (pink). The blue and pink vertical lines indicate the medians in each sample. In panel a, the blue line for the median across all drivers is mostly hidden underneath the pink line for the HOT driver median.

on I-405 while staying as close to the data as possible. While speed and throughput increase in both the GP and HOT lanes after the HOT opening, it is difficult to isolate the effect of pricing alone. Ultimately, the limitations of this long-run analysis motivate the need for our structural analysis to come.

We estimate changes in average speed and throughput by year relative to HOT opening and (GP or HOV/HOT) route. We start by averaging loop speeds and throughputs at the (loop, route, hour, year relative to HOT open) level. The sample is peak hours, southbound 5–11 AM and northbound 1–7 PM, on weekdays from 2011–2019. We then estimate

$$y_{ijht} \sim \text{route}_j \times \text{year}_t + \text{loop}_i + \text{hour}_h \qquad (1)$$

where the two outcomes y_{ijht} are speed and throughput. Figure 2 shows the estimated coefficients on the route-year interactions, where the omitted base level is the GP route in the last (non-calendar) year before the HOT lanes open.

Before the HOT lanes open, the GP and then-HOV lanes appear to be moving along different parts of the backward-bending speed-throughput curve.[19] In the GP lanes, speed and throughput are both decreasing, indicating that these lanes are becoming more and more congested over time. In contrast, in the HOV lanes, speed is decreasing but throughput is *increasing*, suggesting instead

---

[19]Throughput is low both when speed is very low and when speed is very high (Figure C.22b). Recall that throughput is the number of vehicles passing over a given point in space per unit of time. When speed is low, then vehicles take a long time to pass over that point. When speed is high, car-following distances are also high, so that few cars pass over that point per unit of time. Throughput is maximized at an intermediate speed, the max-throughput speed. We can think of the road as being overutilized at lower speeds and underutilized at higher speeds.

Figure 2: Long-run changes in aggregate speed and throughput
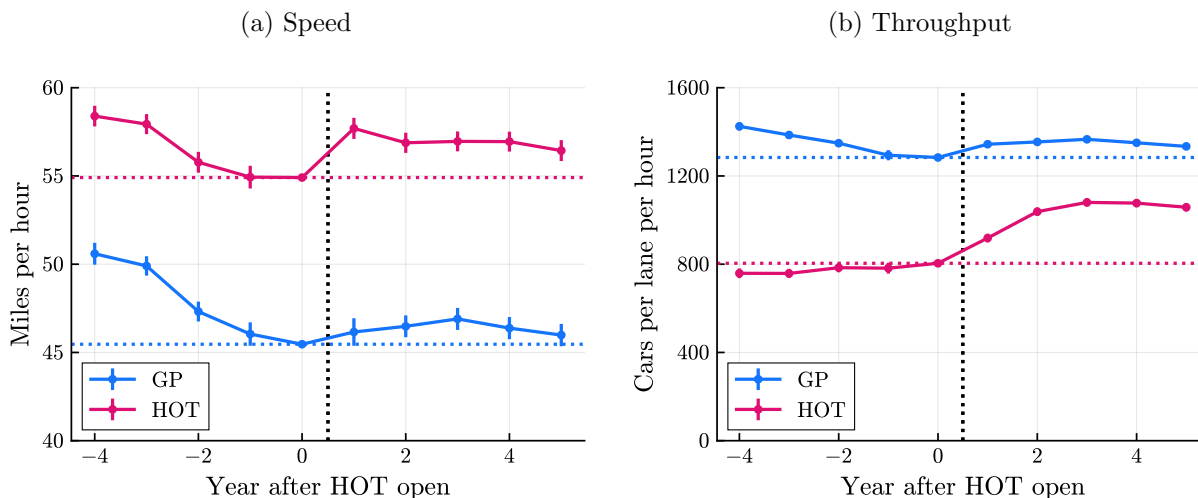
(a) Speed

(b) Throughput

Note: These figures report changes in average speed and throughput by year relative to HOT opening and (GP or HOV/HOT) route. Each point is an estimated coefficient on a route-year interaction in equation (1), with the level normalized to the GP speed or throughput in the last year before the HOT lanes opened. The error bars show 95 percent confidence intervals. The sample is peak hours, southbound 5–11 AM and northbound 1–7 PM, on weekdays from 2011–2019.

that the then-carpool-only lanes are underutilized. HOV speeds average 10 miles per hour faster than GP speeds during this period.

After the HOT lanes open, speed and throughput increase in the short run and begin to decline again in the longer run. The short-run increases are due to the expansion of highway capacity along certain highway segments: a newly constructed HOT lane in September 2015, paired with the introduction of tolling, and a new GP peak-use shoulder lane in April 2017. Estimating equation (1) separately by road segment, we find that HOT throughput increases on all segments (Figure A.4), but speed increases are concentrated on the road segments with the new construction (Figure A.3). Starting about three years after the HOT opening, speed and throughput begin to fall in both lane types. However, the presence of pricing dampens demand increases, so that speed falls more slowly than before tolling was introduced.

This analysis has two main limitations, which together motivate the need for our structural model. First, since the HOT lane opening bundled together several policy changes (detailed at the end of Section 2), it is challenging to satisfactorily separate the effects of the introduction of pricing versus the construction of additional HOT lanes. Even analysis at the road segment level is imperfect, since traffic on a given segment has spillover effects onto traffic on nearby segments. Second, since the 2015 HOT opening falls outside the 2019 sample period of our toll transaction data, this long-run analysis cannot speak to heterogeneous effects by driver characteristics. To overcome these limitations, in the next section, we introduce an equilibrium model of highway travel, which we will use to simulate welfare under counterfactual equilibria.

12

4.2 Suggestive evidence on winners and losers

Next, we document three patterns in the data that suggest that *low-income* drivers may gain the most from I-405 tolling. First, low-income drivers tend to participate in the "longest" markets—i.e., the markets that travel the greatest distances on I-405. Second, the HOT lanes are a better deal in the longer markets. Third, preference heterogeneity along our observable proxies for income appears to be small, so the effect of geography may dominate.

Figure 3 illustrates this first descriptive fact: low-income drivers disproportionately belong to the longest markets. The figure shows how drivers in each tract income quartile are divided among the fourteen southbound markets, ordered visually from shortest (lightest color) to longest (darkest color). While 55 percent of drivers in the bottom quartile participate in the four longest markets, only 30 percent of drivers in the top quartile do. The full-length southbound market (Lynnwood to Bellevue, abbreviated LB in the figure) is particularly overrepresented in the bottom quartile, representing 30 percent of those drivers, compared to no more than 15 percent of drivers in other quartiles. The shortest southbound market (Kirkland to Kirkland), which has no feasible HOT route, has no drivers in the bottom half of the tract income distribution. This pattern is qualitatively similar, though less extreme, when drivers are instead divided into quartiles of car price (Figure A.5).

Figure 3: Driver markets by tract income quartile
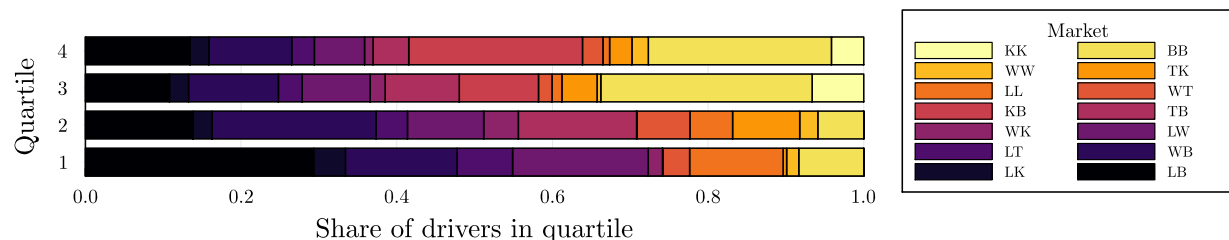
Note: Each underlying observation is a simulated driver in the unconditional sample of potential southbound I-405 drivers. In the legend, markets are ordered from shortest (lightest color) to longest (darkest color).

Not only do the longest markets have disproportionately many low-income drivers, they also offer the best HOT deals. Unsurprisingly, travel times in both the GP and HOT lanes are the greatest in these long markets, with both means and variances of travel times increasing in length of highway traveled (Figure A.6, panels a and b). Since HOT travel times increase more slowly in market length than GP travel times do, HOT time savings are also the greatest in the longest markets. Figure 4a shows that from 7–8 AM, the height of the southbound morning peak, the median time savings are less than two minutes in the four shortest southbound markets with feasible HOT routes, compared to nearly twenty minutes in the full-length market. However, drivers in the longest markets do not pay proportionally higher tolls. In Figure 4b, we see that the price per minute saved *decreases* in market length. The median price per minute saved is $14.65 in one of the shortest markets and only 48 cents in the full-length market.

Figure 4: HOT time savings and price per minute saved by market, 7–8 AM

(a) HOT time savings
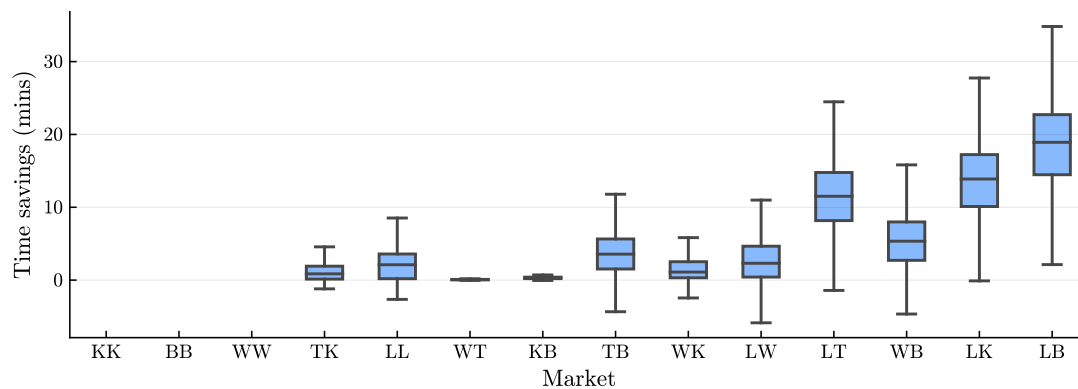
(b) Log of price per minute saved
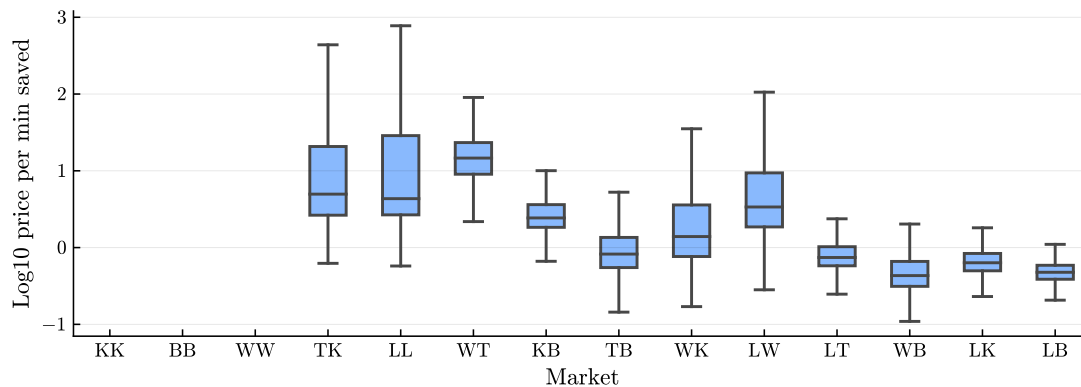
Note: Figures show the distributions of HOT time savings (top panel) and log of price per minute saved (bottom panel) from 7–8 AM in each southbound market. Boxes indicate the 25th percentile, median, and 75th percentile. Lower whiskers extend to the lowest observed data point that is within a distance of 1.5 times the interquartile range (IQR) from the 25th percentile. Likewise, upper whiskers extend to the highest observed data point within 1.5 times the IQR from the 75th percentile. Markets are ordered from shortest to longest. The first three markets have no feasible HOT route.

This across-market concavity of prices in time savings arises from the design of WSDOT's pricing algorithm. There is no clear relationship between market length (hence, travel time or time savings) and price across markets (Figure A.6c). This is because the algorithm sets prices according to contemporaneous traffic conditions, but using worst-case (across road segments traversed by the market) rather than average traffic conditions. The highest prices are found in markets that traverse the highly congested northern half of the corridor, which has only one HOT lane compared to two in the southern half. In these markets, the median price from 7–8 AM is at the $10 ceiling.

Figure 5: Binscatters of HOT trip attributes vs. driver income proxies
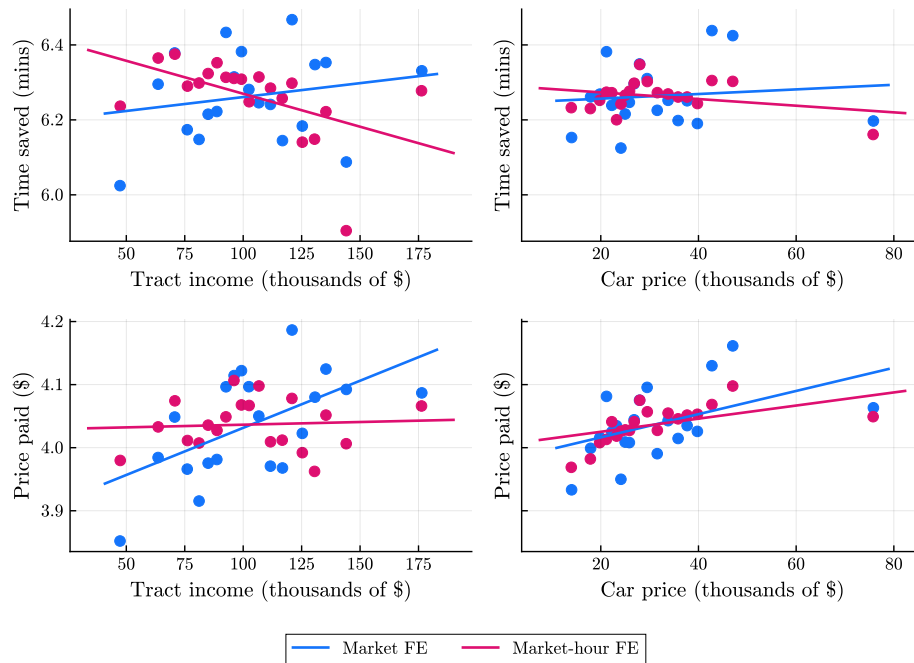
Note: Each underlying observation is a paid peak-hour toll transaction. Southbound peak hours are 5 AM to 11 AM; northbound peak hours are from 1 PM to 7 PM. Binscatters with tract income on the horizontal axis additionally residualize out car price and vice versa.

Differences in aggregate conditions by geography are especially important because preference heterogeneity across observable income measures appears to be modest. Using the toll transaction data, we investigate how the attributes of peak-hour HOT trips vary by the incomes of the drivers taking them. Figure 5 displays binscatters of tract income and car price on the horizontal axes against time saved and price paid on the vertical axes; Table A.2 reports corresponding regression estimates. Comparing HOT transactions within the same market (in blue), trips taken by higher-income drivers tend to have higher time savings and prices. These patterns are partially driven by drivers' choices of *when* to take the toll lanes. Comparing HOT transactions within the same market and hour (in pink), trips by higher-income drivers have *lower* time savings and higher prices, suggesting that they are slightly more willing to pay to avoid travel time. However, the differences are small in magnitude: conditional on the market, hour, and the driver's car price, each $100,000

increase in her tract income is associated with an increase in average time saved by 10.6 seconds and a decrease in average price paid by 0.9 cents.

Taken together, these descriptive patterns suggest that low-income drivers may in fact benefit the most from tolling. Low-income drivers tend to belong to long markets, where the price versus time savings trade-off is particularly favorable. Moreover, these across-market differences may dominate because preference heterogeneity by observable income proxies appears to be small.

5   Model

Motivated by these descriptive facts, we develop an equilibrium model of highway traffic, with the goal of creating a unified framework in which to conduct welfare analyses and simulate alternative policies. In the model, drivers with heterogeneous preferences choose departure times in the first stage and (priced or unpriced) routes in the second stage. Travel times and prices, which adjust in equilibrium to clear the market, are determined respectively by the road technology and the pricing algorithm.

The basic setup of the model is as follows. A **market**, indexed by $m \in \mathcal{M}$, is a highway entry-exit pair. It is possible that a given highway segment will be traversed by drivers from multiple markets, but each driver belongs to exactly one market, which she takes as given. Drivers choose between the unpriced and priced **routes**, indexed by $j \in \{0, 1\}$, in each market. Each day t is independent; time of day h is discretized into five-minute intervals. Travel times, prices, and quantities (of cars on the road) all vary at the (route j, market m, departure time h, day t) level.

5.1   Demand

In the demand model, drivers make sequential departure time and route choices each day. In the first stage, departure times are chosen under imperfect information about traffic conditions. In the second stage, this uncertainty is resolved and routes are chosen under full information. The stages are described below in reverse order. This two-stage structure allows drivers to benefit from tolling in two ways. Drivers who take the priced lanes benefit from realized time savings, which they value more than the price they have to pay. Drivers who don't take the priced lanes ex post nonetheless derive option value from their existence ex ante, as they get to reoptimize in stage 2 after uncertainty about prices and travel times is resolved.

5.1.1   Stage 2: route choice

In the second stage, each driver observes the realizations of price and travel times at her chosen departure time and optimally chooses between the unpriced and priced routes. Drivers have heterogeneous preferences over prices, travel times, and time early or late to their destinations.

Consider a driver i in market m. At this stage, she has already chosen departure time h on day t. Now that she is on the road, she observes the realized price p_{jhmt} and travel time d_{jhmt} (measured in hours) on each route $j \in \{0, 1\}$. Each driver i is endowed with an **ideal arrival**

time $\eta_i \in [0, 24)$. This ideal arrival time combines with her chosen departure time and route to determine her time early or late to her destination.[20]

Driver i chooses a route $j \in \{0, 1\}$ to maximize stage 2 utility

$$u_{i,j|h,m,t} = U_{i,j|h,m,t} + \varepsilon_{i,j|h,m,t} \tag{2a}$$

$$U_{i,j|h,m,t} = \alpha_{jm}^0 + \alpha_i^P \underbrace{p_{jhmt}}_{\text{price}} + \alpha_i^D \underbrace{d_{jhmt}}_{\text{travel time}} + \alpha_i^E \underbrace{(h + d_{jhmt} - \eta_i)_-}_{\text{time early}} + \alpha_i^L \underbrace{(h + d_{jhmt} - \eta_i)_+}_{\text{time late}} \tag{2b}$$

where for any scalar x, we denote the negative part by $x_- = -\min(x, 0)$ and the positive part by $x_+ = \max(x, 0)$. The intercept term α_{jm}^0 captures the baseline attractiveness of route j in market m. The coefficients $(\alpha_i^D, \alpha_i^E, \alpha_i^L)$ are the opportunity costs of time driving, time early, and time late, respectively.[21] The stage 2 utility shocks $(\varepsilon_{i,j|h,m,t})_{j \in \{0,1\}}$ are independently and identically distributed type 1 extreme value.

Given this utility specification, driver i chooses route j with logit choice probability

$$\pi_{i,j|h,m,t} = \frac{\exp U_{i,j|h,m,t}}{\exp U_{i,0|h,m,t} + \exp U_{i,1|h,m,t}}$$

To compute the aggregate market share of route j conditional on departure time h, we integrate over the continuum of drivers who have chosen to depart at h:

$$s_{j|h,m,t} = \int \pi_{i,j|h,m,t} \, dF_{hmt}(i) \tag{3}$$

Finally, the expected value of departure time h *conditional on the realizations* $\mathbf{p}_{hmt} = (p_{jhmt})_{j \in \{0,1\}}$ and $\mathbf{d}_{hmt} = (d_{jhmt})_{j \in \{0,1\}}$ is

$$EU_{ihmt}(\mathbf{p}_{hmt}, \mathbf{d}_{hmt}) = \log\left[\exp U_{i,0|h,m,t}(d_{0hmt}, p_{0hmt}) + \exp U_{i,1|h,m,t}(d_{1hmt}, p_{1hmt})\right]$$

5.1.2 Stage 1: departure time choice

In the first stage, each driver chooses a highway departure time or the non-highway outside option given imperfect information about future prices and travel times. Drivers are forward-looking, comparing their expected values of the different departure times.

Let $\mathcal{H}$ be a discrete set of highway departure times, and let $h = \varnothing$ denote the non-highway outside option. On day t, drivers in market m have common beliefs $G_{hmt}(\mathbf{p}_{hmt}, \mathbf{d}_{hmt})$ about the joint distribution of prices and travel times at each departure time $h \in \mathcal{H}$. There are additionally time-varying shocks ξ_{hmt} to demand for different departure times, which are common across drivers in the market.

---

[20]We only model travel on the highway of interest. Strictly speaking, the chosen departure time h is the time at which the driver enters the highway, rather than the time she leaves home. Similarly, η_i is driver i's ideal time at which she exits the highway, rather than when she arrives at her destination.

[21]Appendix C.1 discusses the microfoundation of this utility specification.

Driver i chooses $h \in \mathcal{H} \cup \{\varnothing\}$ to maximize stage 1 utility u_{ihmt}. The utility from the non-highway outside option is normalized to $u_{i\varnothing mt} = \varepsilon_{i\varnothing mt}$. The utility from departure time h is

$$u_{ihmt} = U_{ihmt} + \xi_{hmt} + \varepsilon_{ihmt} \tag{4a}$$

$$U_{ihmt} = \beta_m^0 + \beta_m^1 \int EU_{ihmt}\left(\mathbf{p}_{hmt}, \mathbf{d}_{hmt}\right) \, \mathrm{d}G_{hmt}\left(\mathbf{p}_{hmt}, \mathbf{d}_{hmt}\right) \tag{4b}$$

The primary component of utility from departure time h is the expected value of departing at h, where the expectation is now additionally taken over drivers' beliefs about prices and travel times. The stage 1 utility shocks $(\varepsilon_{iht})_{h \in \mathcal{H} \cup \{\varnothing\}}$ are independently and identically distributed type 1 extreme value.

Driver i chooses departure time h with logit probability

$$\pi_{ihmt} = \frac{\exp\left\{U_{ihmt} + \xi_{hmt}\right\}}{1 + \sum_{h' \in \mathcal{H}} \exp\left\{U_{ih'mt} + \xi_{h'mt}\right\}}$$

and the overall market share of departure time h is

$$s_{hmt} = \int \pi_{ihmt} \, \mathrm{d}F_m(i) \tag{5}$$

5.2 Road technology

Our road technology model embeds a static relationship between traffic density and traffic speed inside a model of traffic dynamics in space and time. Highway travel is a congestible good: as the density of vehicles on the road increases, vehicles are forced to slow down in order to maintain safe following distances, resulting in lower speeds. By additionally incorporating dynamics, we allow drivers to impose congestion externalities on other drivers traveling not only at the same location and time, but also at other locations and times.[22] The road technology, which maps quantities of drivers on the road to travel times, encodes the technological constraints on the supply of highway travel.

We model the highway as a sequence of **links** indexed by $l \in \{1, \dots, L\}$, each divided into (GP and HOT) routes $j \in \{0, 1\}$. Each link l has length λ_l in miles, and each pair of route j and link l has width κ_{jl}, representing the number of lanes. Time of day is discretized into five-minute intervals of the form $[h, h + \Delta h)$; dates are indexed by t.

The "static" speed-density relationship holds at every (discrete) point in space and time.[23]

---

[22]Previous research on the supply-side costs of congestion, which dates back to Walters (1961), has focused on the estimation of road "supply curves" mapping quantities of vehicles into travel times. Recent papers have estimated these supply curves using rich cross-sectional (Couture, Duranton and Turner, 2018; Akbar et al., 2023) and time series (Mangrum and Molnar, 2018; Yang, Purevjav and Li, 2020; Russo et al., 2021) variation. Combined with a corresponding demand curve mapping travel times into quantities, analysis using the classical Pigou (1920) framework captures the externalities imposed on other drivers traveling at the *same location* and at the *same time* as the focal driver. These "static" congestion costs are amplified in our model of traffic dynamics in space and time.

[23]Appendix C.2 discusses the theoretical relationships between speed, density, and throughput, which we encountered in Section 4.1.

Speed v_{jlht}, in miles per hour, and **density** ρ_{jlht}, in cars per lane-mile, are constant within each route j, link l, five-minute interval h, and date t. Speed is a function of contemporaneous density and a speed shock ψ_{jlht}:

$$v_{jlht} = V\left(\rho_{jlht}, \psi_{jlht}\right)$$

We assume that the mapping $V\left(\cdot, \psi_{jlht}\right)$ is decreasing in density and does not vary across space or time on our highway of interest. The speed shocks are exogenous and capture idiosyncratic deviations from speeds predicted by density alone. For example, drivers on a particular segment at a particular time might be comfortable traveling at higher speeds with shorter following distances.

To close the road technology model, we describe how speeds and densities on *road segments* are related to travel times and quantities in *markets*.[24] As Figure A.7 illustrates, travel times depend on speeds along the entire length of highway traversed and along the entire time interval taken.[25] Similarly, drivers contribute to traffic densities along their entire trajectories. For each driver who departs on route j in market m at time h', let $\underline{h}_{jlmh't}$ and $\overline{h}_{jlmh't}$ denote the (continuous) times at which she respectively enters and exits link l. The density in route j on link l is obtained by summing the cars there from different markets and departure times, then dividing by the total lane-miles:

$$\rho_{jlht} = \frac{1}{\kappa_{jl} \times \lambda_l} \sum_{m \in \mathcal{M}} \sum_{h' \in \mathcal{H}} \underbrace{\frac{\left|[\underline{h}_{jlmh't}, \overline{h}_{jlmh't}) \cap [h, h+\Delta h)\right|}{\Delta h}}_{\substack{\text{fraction of interval } [h, h+\Delta h) \\ \text{in which drivers departing at } h' \\ \text{in market } m \text{ are on link } l}} \times \underbrace{q_{jh'mt}}_{\substack{\text{mass of} \\ \text{departures at } h' \\ \text{in market } m}}$$

5.3 Pricing algorithm

Finally, the pricing algorithm maps (current and potentially past) traffic conditions into prices. In principle, traffic conditions on any link can affect the market m price—for example, if prices in a given market depend on traffic downstream of the segment traversed in that market.

In our empirical setting, the primary pricing algorithm inputs are contemporaneous densities and speeds in both the unpriced ($j = 0$) and priced ($j = 1$) routes. Let $(\boldsymbol{\rho}_{ht}, \mathbf{v}_{ht}) = (\rho_{jlht}, v_{jlht})_{j \in \{0,1\}, l \in \{1,\ldots,L\}}$ denote the vectors of densities and speeds in each route and on each link at time h. The pricing algorithm computes HOT prices using market m-specific functions P_m:

$$p_{0hmt} \equiv 0 \tag{6a}$$

$$p_{1hmt} = P_m\left(\boldsymbol{\rho}_{ht}, \mathbf{v}_{ht}\right) \tag{6b}$$

---

[24]We use a discretized "hydrodynamic" model, in the spirit of the canonical model of Lighthill and Whitham (1955) and Richards (1956), which treats the flow of vehicles analogously to the flow of fluids in physical models.

[25]We model speeds and densities for an additional two hours after the end of the departure time choice set $\mathcal{H}$, assuming there are no departures in these two hours. If a trip still has not concluded by then, we assume the remaining distance is traveled at freeflow speed.

6 Estimation

The demand, road technology, and pricing algorithm components of the equilibrium model are each estimated separately using data from the southbound morning commute in 2019. First, on the demand side, we estimate drivers' heterogeneous preferences for highway travel via the simulated method of moments. Second, for the road technology, we estimate an asymmetric logistic relationship between density and speed. Finally, we approximate the pricing algorithm using market-specific cubic polynomials of GP and HOT travel times into prices. Our data sample throughout is the southbound morning peak, 5–11 AM, in 2019.

6.1 Demand

We estimate the demand model using the simulated method of moments. The primary estimands are drivers' mean preferences for prices, travel times, and time early and late to their destinations—which are identified using plausibly exogenous shifters of price and travel time—and parameters governing preference heterogeneity—which are identified by matching micro moments in the toll transaction data.

Two features of our setting preclude us from directly applying the Berry, Levinsohn and Pakes (1995) method of estimating random-coefficients logit demand models. First, the two stages of the demand model—departure time choice and route choice—are closely linked, requiring joint estimation to account for drivers selecting into departure times based on unobservables. Second, we lack full data on market shares in the two stages: we observe departure time quantities at the *market origin*—not market—level, and we observe HOT but not GP route quantities at the market level. Our approach addresses these data limitations by aggregating demand shocks up to the market origin level and using route *quantities* rather than route *shares* in the moment conditions.

6.1.1 Parameterization

We begin by augmenting the demand model with additional parametric assumptions on drivers' choice sets, preferences, and beliefs. Our empirical model is of the morning commute for drivers traveling southbound on I-405. Drivers choose from departure times $h \in \mathcal{H}$ spaced out every five minutes from 5 AM to 10:55 AM.

In stage 2, drivers have heterogeneous preferences over price and travel time, as well as heterogeneous ideal arrival times. Let $\mathbf{x}_i$ be a vector containing driver i's tract income and car price. Drivers' price and travel time coefficients vary observably with their characteristics and with an unobservable normally distributed component, while their time early and time late coefficients are homogeneous:

$$\begin{bmatrix} \alpha_i^P \\ \alpha_i^D \end{bmatrix} \sim N \left(\begin{bmatrix} \overline{\alpha}^P + \mu^{\alpha,P} \cdot \mathbf{x}_i \\ \overline{\alpha}^D + \mu^{\alpha,D} \cdot \mathbf{x}_i \end{bmatrix}, \ \Sigma^{\alpha,PD} \right) \qquad\qquad \begin{aligned} \alpha_i^E &\equiv \overline{\alpha}^E \\ \alpha_i^L &\equiv \overline{\alpha}^L \end{aligned}$$

Driver characteristics in $\mathbf{x}_i$ are expressed in deviations from the population mean, so that the

common coefficients $\overline{\alpha}$ represent population means. Drivers share common route intercepts α^0_{jm}; without loss of generality, we normalize the GP route intercepts in each market to zero.[26] Ideal arrival times η_i are normally distributed with mean $\overline{\eta}$ calibrated to 8:30 AM and estimated standard deviation σ^η, truncated to the interval between 5 AM and 12 PM.

In stage 1, the primary estimands are the "inside good" intercepts, which control how attractive drivers in each market find I-405 relative to the non-405 outside option. Motivated by data limitations, we assume that demand shocks ξ_{hmt} and inside good intercepts β^0_m vary only at the level of market m's origin:

$$\xi_{hmt} \equiv \xi_{h,\text{orig}(m),t} \qquad\qquad \beta^0_m \equiv \beta^0_{\text{orig}(m)}$$

This dimension reduction is necessary because we observe departure time quantities only at the *market origin*, not market, level. We also fix the stage 1 coefficients β^1_m on the stage 2 expected value at one in all markets. This assumption, which is not without loss of generality, amounts to imposing that the stage 1 and stage 2 logit shocks have the same scale parameter. Importantly, even though the first-stage preference parameters are homogeneous, each driver i's expected value $EU_{ihmt}(\mathbf{p}_{hmt}, \mathbf{d}_{hmt})$ from departure time h still depends on her own specific second-stage preference parameters (α_i, η_i).

For the purpose of estimation, when drivers form beliefs about prices and travel times in stage 1, they observe the quarter of the year (i.e., the season), the day of week, and the presence or absence of morning precipitation.[27] We estimate the joint distribution of prices and travel times in each market, conditional on quarter, day of week, and precipitation, in a first offline step. Appendix D.1 describes this procedure.

Putting it all together, the estimands of the demand model are collected in the vector $\theta = \left(\overline{\alpha}^P, \overline{\alpha}^D, \overline{\alpha}^E, \overline{\alpha}^L, \mu^{\alpha,P}, \mu^{\alpha,D}, \Sigma^{\alpha,PD}, (\alpha^0_{1m})_{m \in \mathcal{M}}, (\beta^0_a)_{a \in \mathcal{A}}, \sigma^\eta\right)$, where market origins are indexed by $a \in \mathcal{A}$.

6.1.2 Price and travel time shifters

Next, we describe three sources of plausibly exogenous variation in prices and travel times, which we use to identify drivers' average preferences for prices and travel times and their scheduling costs. We think of the first two, price rounding and car crashes, as shifting *realized* prices and travel times in stage 2, when the driver is already on the road and is choosing between the GP and HOT routes. We think of the third, precipitation, as shifting drivers' *beliefs* about prices and travel times in stage 1, when she is choosing when to drive.

---

[26] The HOT intercepts capture across-market differences in the convenience of the HOT route. For example, one of the HOT access points is a direct access ramp to the EvergreenHealth medical complex in Totem Lake. GP drivers exiting at Totem Lake must take a different off-ramp, which requires additional driving on suburban roads to get to the medical center.

[27] In counterfactual simulations, we model drivers as having rational expectations—that is, correct beliefs—about prices and travel times.

Price rounding We leverage price discontinuities at the 25-cent rounding thresholds to identify drivers' mean price responsiveness. Unrounded prices on opposite sides of a threshold likely reflect similar underlying market conditions, but are almost-randomly assigned to different realized prices. Figure 6 shows that average HOT transactions drop discontinuously by 1.1 trips per five minutes at the price rounding threshold. The average HOT quantity is 17.2 trips at this level. Away from the threshold, average HOT quantities increase in price, reflecting price endogeneity: higher unrounded prices are associated with higher latent HOT demand.[28]

Figure 6: HOT quantity discontinuity around price rounding threshold
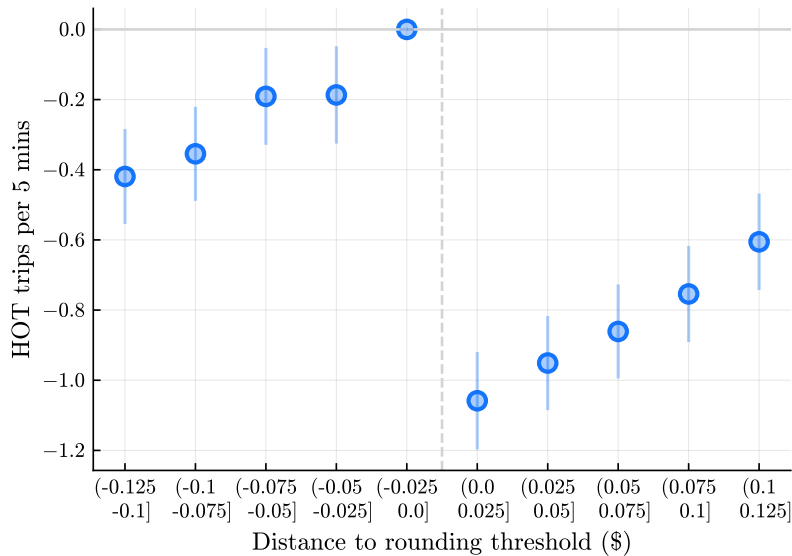
Note: This figure plots coefficients from a regression of HOT quantity on indicators for bins of the unrounded price's distance to the rounding threshold, estimated with threshold and (trip definition, hour) fixed effects. Observations are at the (trip definition, five-minute interval, day) level. Data are subset to southbound 5–11 AM and northbound 1–7 PM.

Figure A.8 shows that realized HOT time savings do not exhibit a similar discontinuity at the price rounding threshold. In principle, the price discontinuity can also affect realized travel times via the discontinuous effect on HOT demand. However, this effect is dampened because travel times depend on demand in both the current five-minute interval (which does respond to this period's price rounding) and future five-minute intervals (which does not). In practice, HOT time savings increase relatively smoothly in the unrounded price.

Moreover, we argue that the discontinuous drop in HOT quantity is not due to drivers updating their beliefs about imperfectly observed travel times. A subset of HOT access points have additional electronic signs displaying estimated travel times in the GP and HOT routes. While the HOT quantity discontinuity persists after restricting to trip definitions which have these signs, again, the

---

[28]We exclude observations where the unrounded price is below the 75-cent price floor or above the $10 price ceiling, as well as observations where rounding our reconstructed unrounded price does not yield the observed rounded price. These dropped observations respectively account for 42 percent, 10 percent, and 2 percent of the original observations.

sign-reported HOT time savings are smooth around the price rounding threshold (Figure A.9).

Car crashes We identify drivers' disutility from travel time using variation from car crashes. Crashes are negative technology shocks, effectively temporarily reducing highway capacity. This in turn results in lower speeds and higher travel times, and in equilibrium, also higher prices. The net effect on HOT time *savings* depends on whether the crash occurred in the GP or HOT lane.

To quantify the effects of GP and HOT crashes on travel times, prices, and HOT quantities, we estimate

$$y_{hmt} \sim \text{GP crash}_{hmt} + \text{HOT crash}_{hmt} + \text{tripDef}_m \times \text{hour}_{ht} \qquad (7)$$

On the right-hand side, we construct indicators for "intersecting" GP and HOT crashes. We define a crash as intersecting a given trip definition at a given time if the crash happened downstream of the trip definition's HOT access point and no more than an hour before the focal time. Observations are at the (five-minute interval h, trip definition m, day t) level.[29]

Table A.3 reports the estimated coefficients. Compared to GP crashes, HOT crashes result in smaller increases in time savings (+0.85 minutes vs. +1.14 minutes) and larger increases in prices (+$1.21 vs. +$0.72). As a result of this unfavorable trade-off, HOT crashes also result in smaller increases in average HOT trips (+0.59 vs. +1.05 trips per five minutes).

Precipitation Finally, we use precipitation, which increases the *variance* of travel times and prices (Figure A.10), to identify drivers' costs of arriving early or late to their destinations. The direction and extent to which drivers shift their departure times in response is informative about the direction and extent of asymmetry in these scheduling costs. If drivers find it much more costly to be late than to be early, then they will choose to depart much earlier on rainy days to avoid potentially incurring those costs.

We find limited departure time responses to this increase in variance, suggesting that drivers have relatively low and slightly asymmetric scheduling costs, with time late slightly more costly than time early. On days with positive precipitation during the morning peak (5–11 AM), the shares of early morning departures increase and the shares of peak-hour departures decrease (Figure A.11a). However, the magnitudes of these differences are extremely small (Figure A.11b).

6.1.3 Estimation and identification

The two stages of the demand model are estimated jointly using the simulated method of moments. The population mean preference parameters are identified using variation from the price and travel time shifters—price rounding, crashes, and precipitation—described above in Section 6.1.2. The preference heterogeneity parameters are identified by matching moments from the toll transaction data, which we first saw in the descriptive evidence in Section 4.2.

---

[29]We exclude the 0.11 percent of observations in which the crash causes WSDOT to close the HOT lanes altogether, which include but are not limited to closures due to very severe crashes.

Joint estimation of the two stages is necessary because of the tight link between the two stages in the demand model. The second-stage parameters are needed to construct the expected value of each departure time in the first-stage utility (4). However, the second stage can't be estimated alone because drivers may select on unobservables into different departure times: for example, drivers who select into departing during peak hours may have lower values of travel time and higher scheduling costs.[30]

We adapt the approach of Berry, Levinsohn and Pakes (1995) to evaluate the moment conditions at each candidate parameter vector θ. At each candidate θ, we first compute $U_{ihmt}(\theta)$, the expected value of departure time h, for each simulated driver, market, and date. Let $\mathcal{M}_a$ denote the set of markets with origin a. For each origin a and date t, we solve a system of nonlinear equations for the vector of demand shocks $\left(\widehat{\xi}_{hat}(\theta)\right)_{h\in\mathcal{H}}$ that rationalize the observed departure time quantities q_{hat}:

$$\underbrace{q_{hat}}_{\substack{\text{departure time } h \\ \text{quantity in} \\ \text{market origin } a}} = \sum_{m\in\mathcal{M}_a} \underbrace{q_{mt}}_{\substack{\text{market} \\ m \text{ size}}} \times \underbrace{\left(\frac{1}{I}\sum_{i=1}^{I} \frac{\exp\left\{U_{ihmt}(\theta) + \xi_{hat}\right\}}{1 + \sum_{h'\in\mathcal{H}}\exp\left\{U_{ih'at}(\theta) + \xi_{h'at}\right\}}\right)}_{\substack{\text{departure time } h \\ \text{share in market } m}} \tag{8}$$

This is done by iterating a BLP-like contraction mapping. Then we use the candidate parameters θ and the implied demand shocks $\widehat{\xi}_{hat}(\theta)$ to predict demand. Note that the vector of unknown demand shocks is at the *market origin* level.[31]

A key challenge is that we only observe departure time quantities at the *market origin*—not market—level, from data on throughputs on highway on-ramps. This data limitation affects our estimation approach in both stages of the demand model. In stage 1, during the BLP inversion (8) of market shares to demand shocks, we are only able to recover demand shocks $\widehat{\xi}_{hat}$ at the market origin a level. In stage 2, since we additionally do not observe GP quantities at the market level, we are unable to compute GP and HOT route *shares* conditional on departure times. Instead, we match moments of HOT route *quantities* in each market, which we do observe in the toll transaction data.

We summarize the following primary moment conditions. In the first stage, we use precipitation as a traditional method of moments instrument, imposing that it is independent of unobserved demand for departure times. In the second stage, we match reduced-form coefficients from regressions of HOT quantity on indicators for price rounding and crashes. To identify the parameters governing preference heterogeneity, we match micro moments of driver characteristics and HOT trip attributes (price paid and time saved) in the toll transaction data. For example, the covariances between a driver's income proxies and the price she pays in her toll transactions are informative about the slope coefficients $\mu^{\alpha,P}$ governing how drivers' price coefficients vary with their incomes.

---

[30]This selection on unobservables can also be seen in equation (3) for the market share of route j in departure time h, where the choice probabilities are integrated over the driver distribution $F_{hmt}(i)$ *conditional on choosing departure time h.*

[31]We approximate the distribution of driver types using $I = 100$ simulated drivers in each market. Driver characteristics are sampled from the unconditional distribution described in Section 3.3.

Appendix D.2 describes the full set of moment conditions in greater detail.

6.1.4 Estimated preferences

Table A.4 reports the estimated demand parameters. The average driver is willing to pay $\overline{\alpha}^D/\overline{\alpha}^P =$ $20.40 to avoid one hour of travel time. Figure 7 shows there is modest heterogeneity in the value of travel time (VOTT), α_i^D/α_i^P. Drivers in the 5th and 95th percentiles of VOTT are willing to pay $17.16 and $22.59, respectively, to avoid an hour of travel time. Drivers' VOTTs are positively correlated with tract incomes and negatively correlated with car prices (Figure A.12). Scheduling costs are relatively low: the average driver is willing to pay $\overline{\alpha}^E/\overline{\alpha}^P =$ $3.53 to avoid being one hour early to her destination and $\overline{\alpha}^L/\overline{\alpha}^P =$ $4.33 to avoid being one hour late. The standard deviation of ideal arrival times, σ^η, is 61 minutes.

Figure 7: Estimated value of travel time distribution
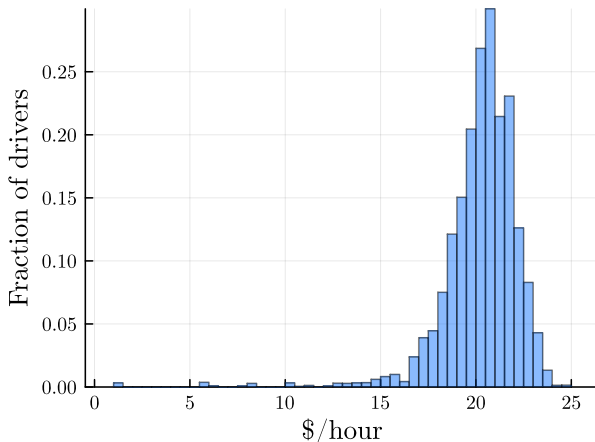
Note: This figure shows the estimated distribution of drivers' value of travel time, α_i^D/α_i^P. Each underlying observation is a simulated driver from the unconditional distribution of driver characteristics.

6.1.5 Discussion

Our empirical demand model comes with important caveats and limitations, which we discuss now in the context of estimation. Section 7 discusses further implications for our simulations of counterfactual equilibria.

First, we assume drivers have perfect information about prices and travel times when they are on the road. In theory, drivers do perfectly observe prices from the electronic signs displayed at each HOT access point (Figure A.1). Moreover, in Section 6.1.2, we tested whether drivers infer HOT time savings from prices in the price rounding regression discontinuity; we found that the discontinuous drop in HOT quantity at the price rounding threshold persists when restricting attention to trip definitions with electronic signs displaying estimated travel times. In practice, however, there may still be drivers who misread the signs or are otherwise confused about how

tolling works on I-405. To the extent that such drivers—who are confused or have less-than-perfect information about traffic conditions once they are already on the road—are common, our revealed preference approach will tend to underestimate drivers' responsiveness to prices and travel times.

We also abstract away from HOV drivers and carpooling choices. This abstraction does not present a problem for estimation if two assumptions hold. First, drivers must take their carpooling status as exogenously given. Second, HOV drivers' preferences over non-price attributes must not be systematically different from those of single-occupancy vehicle (SOV) drivers with the same income proxies. We restrict attention to *paid* HOT trips in both the descriptive analysis and the structural estimation, so those taken by HOV drivers are not included in the paid HOT *quantities* (not market shares) that enter into the demand moment conditions. We will explore relaxing the second assumption above in future versions of this paper.

There are also potentially important dimensions of heterogeneity that are not included or not estimated in the current version of the model. We estimate heterogeneous preferences for price and travel time, which vary observably along income proxies and unobservably. However, we do not estimate heterogeneity in scheduling costs or in the distribution of ideal arrival times.[32] Heterogeneity of this form can in principle be identified using additional data on how departure times (and departure time responses to beliefs shifters like precipitation) vary by driver income proxies.[33] True preferences may also vary *within drivers*, across days. For example, a morning trip to the airport is likely to involve both a different ideal arrival time and higher scheduling costs.[34] If drivers' true ideal arrival times are very different from day to day, then we will tend to underestimate scheduling costs. To identify parameters governing within-driver variation, we would additionally match moments in the toll transaction data on the variability of attributes of HOT trips taken by the same driver.

Finally, we model a medium-run choice problem in which agents decide only when and where they drive. In the long run, households have more margins of adjustment. Workers may rearrange their morning schedules—both earlier and later—in response to traffic conditions, which could lead us to underestimate drivers' true costs of early and late arrivals. Additionally, households may change where they live and work—that is, choose the I-405 market they participate in—in response to changes in transportation policy. Our utility specification includes market-specific HOT intercepts α_{1m}^0, which absorb across-market differences in tastes for HOT lanes. However, our demand estimates may be biased if preferences for price, travel time, and time early and late also vary systematically across markets in a way that is not captured by preference heterogeneity

---

[32] Using survey data from California SR 91 and the National Household Travel Survey, Hall (2020) estimates that drivers' values of travel time are negatively correlated with their scheduling costs. That is, higher-income drivers, who have higher values of time, also have more flexible schedules.

[33] The Puget Sound Regional Council (PSRC) Household Travel Surveys are a promising source of these data, with detailed travel diaries linked to household characteristics. Our preliminary analysis of the 2019 travel diaries shows that the mean morning departure time does not differ substantially by household income in the Seattle metro area. However, the *variance* of morning departure times is higher for low-income households. Though realized departure times are not the same as ideal arrival times, this suggests that the first-order heterogeneity in ideal arrival times is in the variance rather than the mean.

[34] In principle, this within-driver variation is another source of HOT option value.

along income proxies.

6.2 Road technology

To take the road technology model to the data, we estimate the static speed-density relationship specific to I-405. Additionally, in our model of traffic dynamics, we discretize I-405 into $L = 2$ links, corresponding to the congested northern half, which has a single HOT lane, and the more free-flowing southern half, which has two HOT lanes.[35]

The primary empirical object is the speed-density relationship, which is different on every road. It depends on physical factors, including the hilliness of the terrain, the pavement materials and quality, and the presence of medians or other dividers. At each density, we expect speeds to be lower on single-lane, back-country roads than on multilane highways. It also depends on the geometry of the road network: highways, which feature relatively uninterrupted traffic flows, are able to support higher freeflow speeds than dense urban road networks.

We assume the following asymmetric logistic functional form (Wang et al., 2011) mapping traffic density ρ_{jhlt} into traffic speed v_{jhlt}:

$$v_{jlht} = \underline{v} + \frac{\overline{v} - \underline{v}}{\left[1 + \exp\left(\frac{\rho_{jlht} - \rho^*}{\delta_1}\right)\right]^{\delta_2}} + \psi_{jlht} \tag{9}$$

The estimands are the jam speed $\underline{v}$ and the freeflow speed $\overline{v}$ (the lower and upper asymptotes, respectively), a scale parameter δ_1, a parameter δ_2 which controls the degree of asymmetry, and a critical density ρ^* at which traffic transitions from freeflow to congested.

Speed is flat at very low density levels, then decreases convexly in density before finally leveling off.[36] We estimate equation (9) on 2019 loop data from the southbound morning peak. Each observation is a (route j, loop l, five-minute interval h, day t). Figure 8 shows the best-fit curve in pink over a random sample of speeds and densities in blue; Table A.5 reports the estimated parameters. The freeflow speed is $\overline{v} = 63.9$ miles per hour, slightly above the speed limit of 60 miles per hour. Speed begins to fall at the critical density, $\rho^* = 26$ cars per lane-mile, which corresponds to a car-following distance of 2.15 seconds at freeflow speed. The jam speed is $\underline{v} = 7.6$ miles per hour.

6.3 Pricing algorithm

We approximate WSDOT's pricing algorithm using a third-order polynomial of travel times in the unpriced and priced routes, estimated separately for each market. While we observe the true pricing algorithm, its inputs are not easily generated in equilibrium simulations. This approximation is

---

[35]The first, northern link is 8.1 miles long and averages 2.4 GP lanes. The second, southern link is 8.2 miles long and averages 3.2 GP lanes.

[36]The initial flatness is partially an artifact of replacing the top-coded speeds, as the freeflow speeds replacing them are estimated at the less granular (loop, day of week, hour of day) level. However, it is unlikely that average speeds are much higher than 70 miles per hour even when densities are extremely low.

Figure 8: Density vs. speed
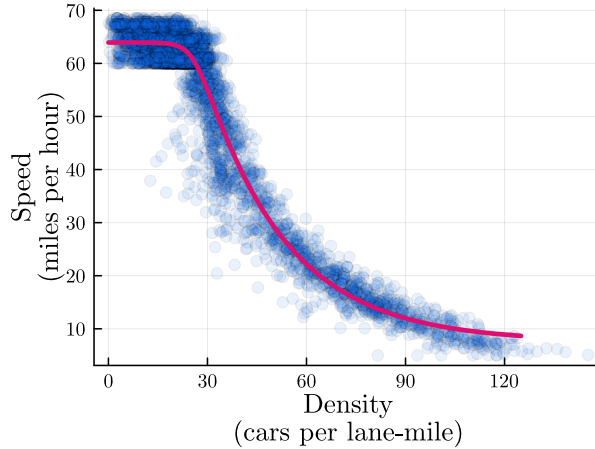
imperfect because realized travel times are a function of traffic conditions not only at the time of departure, but also future traffic conditions. In contrast, the price is set using only information about traffic conditions that is known at departure time. The R^2 values in these regressions range from 0.669 in one of the short markets to 0.933 in one of the long markets.

Prices increase in both GP and HOT travel times, though they increase more steeply in HOT travel times. Figure A.13 shows scatter plots of unrounded prices versus GP travel times, HOT travel times, and HOT time savings in three markets, ordered from longest to shortest. The algorithm-generated unrounded prices in this figure vary between 50 cents and $12; they have not yet been rounded to the nearest 25 cents or constrained to above the 75-cent price floor and below the $10 price ceiling. Longer markets have greater variation in both travel times and prices. In the full-length market, Lynnwood to Bellevue, the freeflow travel time is about 15 minutes and unrounded prices increase steeply to an average of $10 by the time the HOT travel time is about 25 minutes. In the shortest tolled market, Kirkland to Bellevue, the unrounded price is nearly always at its 50-cent lower bound, regardless of travel times in either route.

Since our estimated polynomials are an approximation of the true pricing algorithm, we introduce **price shocks** ϕ_{jhmt} that capture the difference between our predicted unrounded price and the observed unrounded price. We rewrite the pricing equation (6) for our empirical analysis as:

$$p_{0hmt} \equiv 0 \tag{10a}$$

$$p_{1hmt} = \widetilde{P}_m\left(\mathbf{d}_{hmt}\right) + \phi_{1hmt} \tag{10b}$$

where $\widetilde{P}_m$ denotes the estimated polynomial approximation in market m, which maps travel times $\mathbf{d}_{hmt}$ in both the GP and HOT lanes into predicted unrounded prices.

7 Welfare analysis

Armed with our estimated parameters, we now turn to evaluation of status-quo tolling and alternative pricing policies. First, we show that low-income drivers indeed gain the most from status-quo tolling, which benefits drivers across the income distribution but creates winners and losers by geography. Second, we quantify the option value of tolling, which is an important component of welfare gains both in aggregate and for low-income drivers. Finally, we compare status-quo tolling to two types of policy-relevant alternative designs: raising the $10 price ceiling and instituting income-based pricing.

7.1 Aggregate and distributional effects of status-quo tolling

We begin by evaluating the aggregate and distributional effects of the current tolling policy. We find that tolling is aggregate welfare-improving and benefits drivers in nearly all income groups, but particularly drivers in the bottom income quartile. We argue that this is achieved by transferring surplus from relatively high-income drivers in short markets to relatively low-income drivers in long markets.

To do this, we simulate equilibria under two regimes: no toll and status-quo pricing. In the no-toll counterfactual, drivers still choose between two routes in stage 2, but both routes are free.[37] The status quo simulation uses the approximation of the pricing algorithm estimated in Section 6.3. We simulate equilibria using the demand shocks that rationalize the observed departure time shares—that is, by performing the BLP-like inversion (8) at the estimated demand parameters $\widehat{\theta}$.[38] We include utility from all drivers, including those who choose the non-405 outside option; the utility of these non-405 drivers is normalized to zero.

We estimate that status-quo tolling increases aggregate welfare in the southbound morning peak by about $8,500 per day or $2.2 million per year, relative to a world in which the same number of lanes are all free. Table 2 reports aggregate outcomes under the two regimes. Status-quo tolling reduces total driver surplus by about $29,100 per day, but net welfare increases after recycling the $37,600 per day in collected toll revenue. In the status quo, drivers who take I-405 pay an average of 77 cents per day in tolls, but face face lower average travel times (-1.4 minutes) and arrive closer to their ideal arrival times (-2.2 minutes early and -0.9 minutes late) than in the no-toll counterfactual. These time savings primarily arise from reduced mid-morning (6–10 AM) traffic; those drivers largely substitute toward the outside option ($+3.2$ percentage points) rather than to driving in the early or late morning.

Moreover, we find that low-income drivers gain the most from status-quo tolling, which benefits drivers across the income distribution. Table 3 reports changes in the same outcomes split by

---

[37]Maintaining two free routes in the no-toll counterfactual ensures that our welfare differences are not driven by differences in maximizing over one versus two draws from the extreme value distribution.

[38]We also simulate new draws of normally distributed speed shocks ψ_{jlht} and price shocks ϕ_{jhmt}, where the variances of these shocks are estimated as the variances of the residuals of the speed-density equation (9) and the pricing equation (10).

Table 2: Aggregate outcomes under no toll vs. status-quo tolling

	No toll	Status quo	Change
Stage 1 shares			
Outside option	0.459	0.496	+0.036
5–6 AM	0.046	0.045	−0.001
6–7 AM	0.070	0.065	−0.005
7–8 AM	0.121	0.111	−0.010
8–9 AM	0.130	0.118	−0.011
9–10 AM	0.099	0.094	−0.005
10–11 AM	0.075	0.072	−0.004
Stage 2 shares			
GP	1.000	0.721	−0.279
HOT	0.000	0.279	+0.279
Average I-405 outcomes			
Price paid ($)	0.000	0.766	+0.766
Travel time (mins)	14.727	13.350	−1.377
Time early (mins)	24.694	22.453	−2.241
Time late (mins)	12.712	11.851	−0.861
Welfare			
Driver surplus ($K/day)	236.313	207.250	−29.062
Toll revenue ($K/day)	0.000	37.590	+37.590
Net welfare ($K/day)	236.313	244.841	+8.528

Note: The stage 2 shares and average I-405 outcomes are conditional on choosing I-405—that is, conditional on *not* choosing the non-405 outside option. The welfare values sum across all drivers, including those who choose the outside option. Average time early averages zeros (if the driver is late) and positive values (if the driver is early); the same is true for average time late.

tract income quartile. After uniform redistribution of toll revenues, net welfare in the bottom tract income quartile increases by $5,000 per day, or about 59 percent of the $8,500 aggregate welfare gain. Conditional on choosing I-405, drivers in the bottom quartile pay the highest average tolls ($1.30 per day) but save the most in travel time (−2.3 minutes), time early (−4.5 minutes), and time late (−1.5 minutes) relative to in the unpriced equilibrium; they also take the HOT lanes at the highest rate (38 percent). Drivers in the top three quartiles experience smaller reductions in travel time, time early, and time late. While status-quo tolling incentivizes substitution to the outside option in the top three tract income quartiles, drivers in the bottom quartile take I-405 slightly more frequently under tolling. Table A.7 documents that, consistent with the suggestive evidence presented in Section 4.2, bottom-quartile drivers face the highest travel times and incur the greatest scheduling costs in the unpriced equilibrium. Status-quo tolling dampens but does not eliminate these patterns.[39]

---

[39]The results in this paragraph are all qualitatively similar when drivers are instead split by car price quartile (Table A.6).

Table 3: No toll vs. status-quo tolling differences by tract income quartile

	Q1	Q2	Q3	Q4
Stage 1 shares				
Outside option	+0.002	+0.015	+0.053	+0.072
5–6 AM	−0.001	0.000	−0.001	−0.002
6–7 AM	−0.004	−0.004	−0.006	−0.008
7–8 AM	+0.001	−0.004	−0.014	−0.021
8–9 AM	−0.002	−0.003	−0.018	−0.021
9–10 AM	+0.002	−0.001	−0.007	−0.013
10–11 AM	+0.002	−0.003	−0.007	−0.007
Stage 2 shares				
GP	−0.379	−0.299	−0.237	−0.210
HOT	+0.379	+0.299	+0.237	+0.210
Average I-405 outcomes				
Price paid ($)	+1.300	+0.756	+0.585	+0.460
Travel time (mins)	−2.286	−1.422	−1.032	−0.836
Time early (mins)	−4.487	−1.987	−1.593	−1.055
Time late (mins)	−1.484	−0.864	−0.622	−0.519
Welfare				
Driver surplus ($K/day)	−3.995	−7.419	−8.772	−8.877
Toll revenue ($K/day)	+12.252	+9.898	+8.099	+7.342
Net welfare ($K/day)	+4.998	+1.902	+0.601	+1.027

Note: This table reports differences in outcomes between the no-toll counterfactual and status-quo tolling (corresponding to the third column of Table 2) separately by quartiles of tract income. The quartiles generate different amounts of toll revenue, but the net welfare values assume uniform redistribution of that toll revenue across all drivers.

The large gains for low-income drivers are largely driven by geography, a dimension along which status-quo tolling does create losers. We documented in Section 4.2 that low-income drivers disproportionately live in the longest markets. Figure 9 shows that tolling transfers surplus from relatively high-income drivers in short markets to relatively low-income drivers in long markets. The magnitudes of these transfers are substantial, ranging from a $3,200 per day loss in one of the shortest markets to a $8,500 per day gain in the full-length market—almost identical to the aggregate welfare gains obtained by summing across all markets. Drivers in short markets face low HOT time savings at relatively high prices, so they frequently experience the downsides of increased GP congestion and rarely enjoy the faster HOT lanes. This is especially the case in the three shortest southbound markets, for which there is no feasible HOT route altogether.

Discussion Before proceeding, we describe the key assumptions and limitations of our equilibrium simulations and discuss how they affect the interpretation of these results.

The first assumption, which has the most bite, is that we abstract away from HOV drivers.

Figure 9: Aggregate welfare changes by market
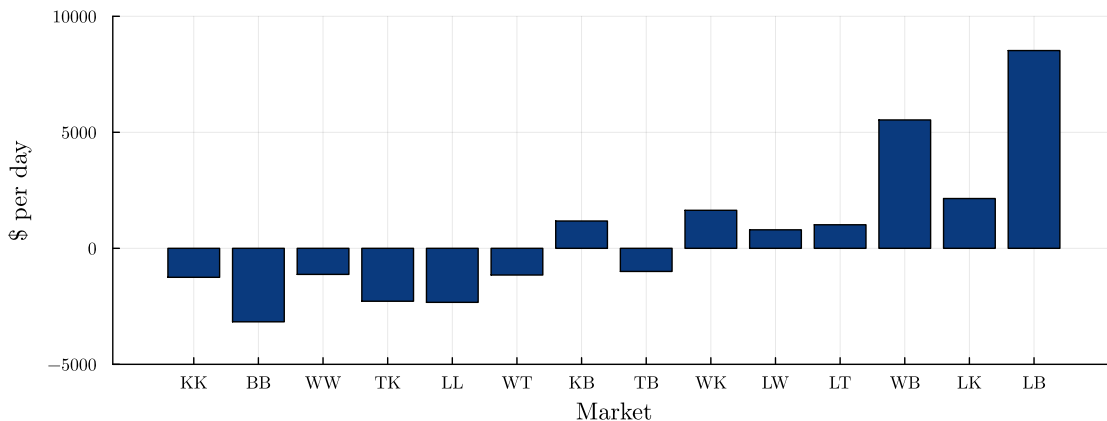
Note: This figure shows differences in aggregate welfare between the no-toll counterfactual and the status quo, separately for each southbound market. Markets are ordered from shortest to longest.

When we simulate the status-quo equilibrium, the lack of HOV trips means that simulated travel times, and hence simulated prices, are low in the HOT lanes. Our current estimates therefore overestimate the time saved in the HOT lanes and underestimate the price paid, though we also underestimate the total revenue generated from tolling, which is redistributed uniformly among potential I-405 drivers in our welfare calculations.[40]

Additionally, our simulations hold the value of the non-405 outside option fixed across counterfactual equilibria. In reality, the welfare gains from tolling on I-405 could be partially offset by welfare losses from increased congestion on substitute roads and highways. However, these offsetting effects are unlikely to change our finding that low-income drivers gain the most from status-quo tolling, because lower-income drivers tend to participate in markets with lower shares of the outside option. Table A.4 shows that the estimated inside good intercepts, which control the baseline attractiveness of I-405 relative to the outside option, are higher for more northern—and relatively lower-income—market origins.

Finally, in these counterfactuals, drivers choose new departure times and routes, but they do not make long-run adjustments like rearranging their morning schedules or changing where they live or work. Our results should therefore be interpreted through the lens of this medium run, in which households only make decisions about driving. Endogenizing the longer-run responses would likely amplify the effects of introducing congestion pricing.[41]

---

[40]The underestimate of toll revenue is due almost entirely to underestimating the intensive margin (the average price paid) rather than the extensive margin (the number of paid trips).

[41]Recent empirical papers which model joint commuting, employment, and residential choice find that congestion pricing incentivizes households to reduce their commuting distances and substitute away from commutes through highly congested areas (Herzog, 2020; Barwick et al., 2021).

7.2 Ex ante and option values of tolling

Next, we decompose the welfare changes from status-quo tolling into two channels, the ex ante value and the option value of tolling. Since drivers make departure time choices in the first stage of the demand model under imperfect information, they derive option value from being able to reoptimize their route choices in the second stage, after the uncertainty is resolved. The ex ante value captures the value of tolling in the absence of this ability to reoptimize. Since it also accounts for the re-equilibration of GP travel times after the introduction of tolling, it can be positive or negative, while the option value is positive by construction.

To quantify these two welfare channels, we compute welfare in an intermediate scenario with full first-stage commitment. In this intermediate step, drivers continue to face status-quo prices and travel times, but they must commit to both their departure time and route choices in stage 1, before they observe the realizations of prices and travel times in stage 2. A driver's option value from tolling is the difference between her status-quo welfare and her welfare under commitment. The ex ante value is the remaining difference between welfare under commitment and welfare in the no-toll counterfactual.[42]

Table 4: Ex ante and option values of tolling in aggregate and by tract income quartile

	Aggregate	By tract income quartile			
		Q1	Q2	Q3	Q4
Ex ante value (\$K/day)	+1.598	+2.933	−0.520	−0.619	−0.195
Option value (\$K/day)	+6.930	+2.065	+2.423	+1.220	+1.222

We find that the option value is an important driver of aggregate welfare gains from status-quo tolling: we estimate its aggregate value to be about \$6,900 per day, accounting for 81 percent of the overall gains of \$8,500 per day. The ex ante value makes up the remaining \$1,600 per day in aggregate. Table 4 reports ex ante and option values in aggregate and by tract income quartile; Figure A.14 shows how this decomposition varies across markets. Drivers in the bottom half of tract incomes have higher option values of tolling (+\$2,100 to +\$2,400 versus about +\$1,200), but the option value is largely similar in dollar terms across markets. On the other hand, there is substantial variation in ex ante values across both income quartiles and markets. Drivers in the bottom tract income quartile have the highest ex ante values of tolling (+\$2,900 per day), while the ex ante values of drivers in the top three quartiles are negative. These differences are again driven by geography: the ex ante value is large and negative in relatively high-income short markets and large and positive in relatively low-income long markets.

To summarize, low-income drivers benefit from *both* high ex ante values and high option values of tolling. They disproportionately participate in long markets, where HOT time savings are high

---

[42]Note that this is not a counterfactual *equilibrium*: we do not recompute equilibrium prices and travel times in this commitment world. Instead, the thought experiment is that one driver at a time is forced to go about her highway travel under full first-stage commitment. Since each driver is small and can't affect aggregate traffic conditions by herself, the equilibrium prices and travel times are unchanged.

and prices are relatively low, resulting in high ex ante values. Moreover, since low-income drivers are less willing to pay to avoid travel time, they are more marginal when deciding between the GP and HOT routes. As a result, they benefit more from being able to make that choice under full information about the trade-offs.

7.3 Alternative pricing policies

Finally, we compare status-quo tolling to two types of alternative pricing policies that are currently under consideration by WSDOT:

1. **Higher price ceiling:** As of October 2023, WSDOT is considering raising the $10 price ceiling to $15 or $18 in order to control increasing HOT congestion and make up revenue lost during the pandemic (Lindblom, 2023). We simulate a counterfactual equilibrium in which the price ceiling is $12. We consider this more modest increase in order to avoid extrapolating our pricing algorithm approximation beyond the historical unrounded prices, which are capped at $12.

2. **Low-income discount:** In 2021, the Washington State Transportation Commission (WSTC) conducted a study exploring options for a potential Low-Income Tolling Program (Washington State Transportation Commission, 2021). We evaluate two versions of income-based tolling that they consider in that study: a 50 percent discount and a $2 per trip discount.[43] In our counterfactual simulations, drivers in the bottom quartile of tract income are eligible for this discount.[44]

Table 5 reports how (departure time and route) market shares, prices and travel times, and welfare change under these alternative policies relative to the status quo. We report both aggregate changes and changes for drivers in the bottom tract income quartile, who are targeted for the low-income discounts.

Raising the price ceiling to $12 is welfare-improving both in aggregate and across all quartiles of tract income and car price. The greater flexibility afforded by this higher ceiling allows the pricing algorithm to more efficiently manage peak congestion without substantially increasing average prices paid. Under the higher ceiling, drivers pay very slightly higher average tolls (+3 cents per day) and face a small *increase* in time late (+0.011 minutes), but obtain modest reductions in average travel time (−0.101 minutes) and time early (−0.144 minutes). Aggregate welfare gains are $1,700 per day higher. Drivers in the bottom tract income quartile, who drive the longest

---

[43]The WSTC study considers three additional options: a fixed toll credit per month, a fixed number of free toll trips per month, and a lower price ceiling for low-income drivers only. In April 2023, the San Francisco Bay Area Metropolitan Transportation Commission began piloting its own income-based tolling program, Express Lanes START, which offers low-income drivers a 50 percent discount (Metropolitan Transportation Commission, 2023).

[44]The most common eligibility criterion, both considered in the WSTC study and actually in place in the Bay Area, is that the driver's household income is less than 200 percent of the federal poverty line. We do not directly observe household incomes, so we choose this 25th percentile cutoff for tract median household income based on ACS estimates that 23.3 percent of Washington State households had incomes below 200 percent of the federal poverty line in 2019 (KFF, 2023).

Table 5: Changes in outcomes under alternative policies

	Raise price ceiling to \$12		Low-income 50% discount		Low-income \$2 discount	
	Agg	Q1	Agg	Q1	Agg	Q1
Stage 1 shares						
Outside option	0.000	−0.001	+0.001	−0.030	−0.001	−0.035
5–6 AM	0.000	0.000	+0.000	+0.001	+0.000	+0.002
6–7 AM	+0.000	+0.000	+0.001	+0.007	+0.001	+0.006
7–8 AM	+0.000	+0.001	+0.000	+0.010	+0.002	+0.011
8–9 AM	0.000	0.000	−0.001	+0.011	−0.001	+0.011
9–10 AM	0.000	+0.000	−0.001	+0.003	−0.001	+0.004
10–11 AM	+0.000	+0.001	−0.001	0.000	0.000	+0.002
Stage 2 shares						
GP	−0.002	−0.001	+0.010	−0.034	−0.003	−0.049
HOT	+0.002	+0.001	−0.010	+0.034	+0.003	+0.049
Average I-405 outcomes						
Price paid (\$)	+0.031	+0.052	−0.311	−0.448	−0.222	−0.336
Travel time (mins)	−0.101	−0.179	+0.650	+1.221	+0.415	+0.713
Time early (mins)	−0.144	−0.268	+0.954	−0.185	+0.591	−0.261
Time late (mins)	+0.011	+0.020	+0.391	−0.247	+0.378	+0.092
Welfare						
Driver surplus (\$K/day)	+0.346	+0.259	+0.523	+1.442	+2.445	+2.526
Toll revenue (\$K/day)	+1.325	+0.425	−10.534	−1.512	−8.891	−1.506
Net welfare (\$K/day)	+1.671	+0.576	−10.011	−1.079	−6.446	+0.399

Note: This table reports changes in outcomes under alternative policies relative to the status quo, computed both in aggregate and for the bottom tract income quartile. In both cases (aggregate and Q1), we sum market shares and welfare across all drivers. The net welfare values assume uniform redistribution of toll revenue across all drivers.

distances on I-405, also gain relative to the status quo. They also pay very slightly higher average tolls (+5 cents per day) and face a small increase in time late (+0.020 minutes), but enjoy larger reductions in average travel time (−0.179 minutes) and time early (−0.268 minutes).

While both versions of the low-income discount increase welfare relative to the completely unpriced equilibrium, they are worse than the status quo, in aggregate and even for the targeted drivers in the bottom tract income quartile. Under these discounts, low-income drivers still impose the same externality on other drivers as their high-income peers do, but they pay less, incentivizing them to take the toll lanes inefficiently often. Under the 50 percent discount, the HOT share for bottom-quartile drivers increases (+3.4 percentage points); these HOT trips also shift toward the most congested part of the morning, when prices and HOT time savings are high and the 50 percent discount is large in absolute terms. Under the \$2 per trip discount, the HOT share for bottom-quartile drivers increases even more (+4.9 percentage points), but these HOT trips shift toward the early and late morning, when congestion is relatively low. Both forms of income-based

tolling result in higher travel times and lower toll revenue to be redistributed.

While we have so far evaluated realistic modifications to status-quo tolling, a natural next question is how all of these pricing schemes compare to first-best pricing of all of the lanes. This analysis, which is computationally intensive, is ongoing. The additional gains from first-best pricing are likely to be large. The current pricing algorithm, whose objective is to keep HOT speeds above 45 miles per hour, deviates from optimal pricing in several key ways. First and most obviously, only a subset of the highway lanes are tolled, so congestion externalities remain unpriced in the GP lanes. Second, while the second-best pricing scheme would seek to maximize the welfare of drivers in all lanes by setting prices in the toll lanes, the existing algorithm cares only about drivers in the toll lanes. Third, guaranteeing a minimum speed need not be welfare-maximizing even for drivers in the toll lanes.

Nonetheless, though it may be far from first-best in aggregate terms, status-quo tolling on I-405 demonstrates the feasibility of implementing highway congestion pricing in a way that benefits drivers across the income distribution.

8 Conclusion

This paper studies the aggregate and distributional effects of highway congestion pricing, which is often thought to be regressive. We show that in our empirical setting, it is actually *low-income* drivers who gain the most from status-quo tolling. Low-income drivers have high ex ante values of tolling because they drive the longest distances on our focal highway, in markets where time savings in the toll lanes are high and prices are relatively low. At the same time, these drivers, who are less willing or able to pay to avoid travel time, also derive considerable value from having the option but not the obligation to pay for speed.

While discussion of the potential regressiveness of congestion pricing often focuses on how low-income drivers must spend a greater fraction of their incomes on congestion fees, our results highlight that low-income drivers often also bear the greatest costs of unpriced traffic congestion. In many urban and suburban areas around the United States, high housing costs have pushed low-income households to live increasingly far away from job centers, where they must both drive longer distances to work and drive more often due to limited public transit. While in our setting, low-income drivers do pay higher tolls on average than other drivers—even in dollars, not only as a share of their incomes—we nonetheless find that congestion pricing is *less regressive* than completely unpriced congestion.

Our findings also point to the importance of the option value channel, which is relevant in potentially many other settings. The trade-off between paying with time versus paying with money is not unique to highway tolling. Both Uber and Lyft offer riders the choice between a more expensive trip with a faster pickup and a cheaper trip that requires additional waiting time. At airports, travelers can pay for CLEAR Plus in order to skip long airport security lines. Even at Disney theme parks, the Genie+ program allows paying customers access separate, faster "Lightning

Lanes" for rides. In any setting where there is uncertainty about future prices or travel or wait times, even market participants who are unlikely to pay ex ante may still value having the option to pay if it turns out to be necessary ex post.

References

Akbar, Prottoy A., Victor Couture, Gilles Duranton, and Adam Storeygard. 2023. "The fast, the slow, and the congested: Urban transportation in rich and poor countries." https://www.nber.org/papers/w31642.

Anderson, Michael L., and Lucas W. Davis. 2020. "An empirical test of hypercongestion in highway bottlenecks." *Journal of Public Economics*, 187: 104–197.

Arnott, Richard, André de Palma, and Robin Lindsey. 1994. "The welfare effects of congestion tolls with heterogeneous commuters." *Journal of Transport Economics and Policy*, 28(2): 139–161.

Arora, Kashish, Fanyin Zheng, and Karan Girotra. 2020. "Private vs. pooled transportation: Customer preference and congestion management." https://papers.ssrn.com/abstract=3701056.

Astor, Maggie. 2017. "A $40 toll to drive 10 miles? It happened on Virginia's I-66." *The New York Times*.

Barwick, Panle Jia, Shanjun Li, Andrew R. Waxman, Jing Wu, and Tianli Xia. 2021. "Efficiency and equity impacts of urban transportation policies with equilibrium sorting." https://www.nber.org/papers/w29012.

Belenky, Peter. 2011. "Revised departmental guidance on valuation of travel time in economic analysis." U.S. Department of Transportation.

Bento, Antonio, Kevin Roth, and Andrew Waxman. 2020. "Avoiding traffic congestion externalities? The value of urgency." https://www.nber.org/papers/w26956.

Berry, Steven T., James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica*, 63(4): 841–890.

Börjesson, Maria, Jonas Eliasson, Muriel B. Hugosson, and Karin Brundell-Freij. 2012. "The Stockholm congestion charges-5 years on. Effects, acceptability and lessons learnt." *Transport Policy*, 20: 1–12.

Brent, Daniel A., and Austin Gross. 2018. "Dynamic road pricing and the value of time and reliability." *Journal of Regional Science*, 58(2): 330–349.

Buchholz, Nicholas, Laura Doval, Jakub Kastl, Filip Matějka, and Tobias Salz. 2020. "The value of time: Evidence from auctioned cab rides." http://www.nber.org/papers/w27087.pdf.

Bureau of Labor Statistics. 2020. "Occupational employment and wages in Seattle-Tacoma-Bellevue — May 2019."

Castillo, Juan Camilo. 2019. "Who benefits from surge pricing?" https://web.stanford.edu/~jccast/JMP_Castillo.pdf.

Castillo, Juan Camilo, Daniel T. Knoepfle, and E. Glen Weyl. 2023. "Matching and pricing in ride hailing: Wild goose chases and how to solve them." https://papers.ssrn.com/abstract=2890666.

Cho, Sungjin, Gong Lee, John Rust, and Mengkai Yu. 2018. "Optimal dynamic hotel pricing." https://economics.sas.upenn.edu/index.php/system/files/2018-04/hp_final_update.pdf.

Couture, Victor, Gilles Duranton, and Matthew A. Turner. 2018. "Speed." *Review of Economics and Statistics*, 100(4): 725–739.

D'Haultfœuille, Xavier, Ao Wang, Philippe Février, and Lionel Wilner. 2022. "Estimating the gains (and losses) of revenue management." https://arxiv.org/abs/2206.04424v3.

Glaeser, Edward L., Caitlin S. Gorback, and James M. Poterba. 2022. "How regressive are mobility-related user fees and gasoline taxes?" https://www.nber.org/papers/w30746.

Goldszmidt, Ariel, John A. List, Ian Muir, Robert D. Metcalfe, V. Kerry Smith, and Jenny Wang. 2020. "The value of time in the United States: Estimates from nationwide natural field experiments." https://papers.ssrn.com/abstract=3748629.

Hallenbeck, Mark, Shirley Leung, Cory McCartan, CJ Robinson, Kiana Roshan Zamir, and Vaughn Iverson. 2019. "I-405 Express Toll Lanes analysis: Usage, benefits, and equity." https://depts.washington.edu/trac/research-news/freeway-and-arterial-management/i-405-express-toll-lanes-analysis-usage-benefits-and-equity/.

Hall, Fred L. 2005. "Traffic stream characteristics." In *Revised Monograph on Traffic Flow Theory*. Ed. Nathan Garner, Carroll J. Messer and Ajay K. Rathi, Chapter 2, 1–32. Federal Highway Administration.

Hall, Jonathan 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.

Hall, Jonathan D. 2020. "Can tolling help everyone? Estimating the aggregate and distributional consequences of congestion pricing." *Journal of the European Economic Association*, 19(1): 441–474.

Herzog, Ian. 2020. "The city-wide effects of tolling downtown drivers: Evidence from London's congestion charge." https://www.dropbox.com/s/j9w8g8qwqif1ide/HerzogLondonJMP.pdf.

KFF. 2023. "Distribution of total population by federal poverty level." https://www.kff.org/other/state-indicator/distribution-by-fpl/?currentTimeframe=1.

Kreindler, Gabriel. 2023. "Peak-hour road congestion pricing: Experimental evidence and equilibrium implications." https://www.nber.org/papers/w30903.

Lam, Terence C., and Kenneth A. Small. 2001. "The value of time and reliability: measurement from a value pricing experiment." *Transportation Research Part E: Logistics and Transportation Review*, 37(2-3): 231–251.

Leape, Jonathan. 2006. "The London congestion charge." *Journal of Economic Perspectives*, 20(4): 157–176.

Ley, Ana. 2023. "Congestion pricing plan in New York City clears final federal hurdle." *New York Times*.

Lighthill, Michael James, and Gerald Beresford Whitham. 1955. "On kinematic waves II. A theory of traffic flow on long crowded roads." *Proceedings of the Royal Society of London Series A*, 229(1178): 317–345.

Lindblom, Mike. 2023. "A $15 toll? How about $18? WSDOT may blow the lid off I-405 express-lane prices." *The Seattle Times*.

Mangrum, Daniel, and Alejandro Molnar. 2018. "The marginal congestion of a taxi in New York City." https://www.danielmangrum.com/docs/Boro_current.pdf.

Martin, Leslie A., and Sam Thornton. 2018. "The margins of response to road use prices."

Mattia, Andrea. 2022. "The value of time: Evidence from traffic congestion and Express Lanes." https://www.dropbox.com/s/jp78r9agui9ll73/Mattia_JMP_VOT.pdf.

Metropolitan Transportation Commission. 2023. "I-880 Express Lanes now more affordable for low-income travelers." https://mtc.ca.gov/news/i-880-express-lanes-now-more-affordable-low-income-travelers.

Phang, Sock-Yong, and Rex S. Toh. 2004. "Road congestion pricing in Singapore: 1975 to 2003." *Transportation Journal*, 43(2): 16–25.

Pigou, Arthur Cecil. 1920. *The Economics of Welfare.* Palgrave MacMillan.

Richards, Paul I. 1956. "Shock waves on the highway." *Operations Research*, 4(1): 42–51.

Rosaia, Nicola. 2020. "Competing platforms and transport equilibrium: Evidence from New York City." https://scholar.harvard.edu/files/rosaia/files/draft.pdf.

Rosendorf, Linda. 2018. "Virginia's 'Lexus lanes' deserve their nickname." *The Washington Post*.

Russo, Antonio, Martin W. Adler, Federica Liberini, and Jos N. van Ommeren. 2021. "Welfare losses of road congestion: Evidence from Rome." *Regional Science and Urban Economics*, 89: 103692.

Samdahl, Donald, Myron Swisher, Jennifer Symoun, and Will Lisska. 2013. "Congestion pricing: A primer on institutional issues." Federal Highway Administration.

Small, Kenneth A. 2012. "Valuation of travel time." *Economics of Transportation*, 1(1-2): 2–14.

Small, Kenneth A., and Jia Yan. 2001. "The value of 'value pricing' of roads: Second-best pricing and product differentiation." *Journal of Urban Economics*, 49(2): 310–336.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. "Uncovering the distribution of motorists' preferences for travel time and reliability." *Econometrica*, 73(4): 1367–1382.

Sound Transit. 2023. "Paying for regional transit." https://www.soundtransit.org/get-to-know-us/paying-regional-transit/car-tab-motor-vehicle-excise-tax-mvet.

Su, Yichen. 2019. "The rising value of time and the origin of urban gentrification." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3494191.

Tarduno, Matthew. 2022. "For whom the bridge tolls: Congestion, air pollution, and second-best road pricing." https://matt-tarduno.github.io/jmp/.

Verhoef, Erik T., and Kenneth A. Small. 2004. "Product differentiation on roads: Constrained congestion pricing with heterogeneous users." *Journal of Transport Economics and Policy*, 38(1): 127–156.

Vickrey, William S. 1969. "Congestion theory and transport investment." *American Economic Review: Papers & Proceedings*, 59(2): 251–260.

Walters, Alan A. 1961. "The theory and measurement of private and social cost of highway congestion." *Econometrica*, 29(4): 676.

Wang, Haizhong, Jia Li, Qian Yong Chen, and Daiheng Ni. 2011. "Logistic modeling of the equilibrium speed–density relationship." *Transportation Research Part A: Policy and Practice*, 45(6): 554–566.

Washington State Department of Transportation. 2023. "Ramp meters." https://wsdot.wa.gov/travel/operations-services/ramp-meters.

Washington State Transportation Commission. 2021. "Low-income toll program study for I-405 & SR 167 Express Toll Lanes."

Williams, Kevin R. 2021. "The welfare effects of dynamic pricing: Evidence from airline markets." https://www.nber.org/papers/w28989.

Wood, Nick, Vivek Gupta, James P. Cardenas, Jinuk Hwang, and Deepak Raghunathan. 2021. "National inventory of specialty lanes and highways: Technical report." Federal Highway Administration.

Yang, Jun, Avralt Od Purevjav, and Shanjun Li. 2020. "The marginal cost of traffic congestion and road pricing: Evidence from a natural experiment in Beijing." *American Economic Journal: Economic Policy*, 12(1): 418–53.

A   Additional figures and tables

## A.1   Setting and data

Figure A.1: Example HOT access point

Note: This photograph shows an HOT access point on southbound I-405 just after the State Route 527 interchange. There are two GP lanes on the right and one HOT lane on the left. The GP and HOT lanes are separated at the bottom of the frame by double white lines; the access point is toward the top of the frame, where the double white lines become a single dashed line where drivers may legally cross. Above this access point, an electronic sign displays prices for travel to each downstream HOT access point (i.e., prices for each trip definition that begins at this access point). Source: Ken Lambert, *The Seattle Times*, https://www.seattletimes.com/seattle-news/transportation/a-15-toll-how-about-18-wsdot-may-blow-the-lid-off-i-405-express-lane-prices/

Table A.1: Southbound median traffic conditions by market

Market	Length (miles)	GP travel time (mins)		HOT travel time (mins)		Time savings (mins)		Price ($)	
		Peak	Off-peak	Peak	Off-peak	Peak	Off-peak	Peak	Off-peak
KK	0.7	0.7	0.7	–	–	–	–	–	–
BB	1.1	1.2	1.5	–	–	–	–	–	–
WW	1.9	5.2	1.9	–	–	–	–	–	–
TK	3.2	4.3	3.2	4.0	3.2	0.1	0.0	1.75	0.75
LL	3.3	4.0	3.3	3.8	3.3	0.0	0.0	5.25	0.75
WT	4.4	8.5	4.5	8.3	4.5	0.0	0.0	0.75	0.75
KB	4.4	5.2	5.0	4.8	4.4	0.3	0.4	0.75	0.75
TB 1	6.9	9.1	7.6	7.8	6.9	0.9	0.6	0.75	0.75
TB 2	6.9	9.1	7.6	6.9	6.9	2.1	0.7	1.75	0.75
WK	7.0	12.2	7.1	11.5	7.1	0.3	0.0	0.75	0.75
LW	7.4	14.6	7.6	13.0	7.6	0.8	0.0	5.25	0.75
LT	10.0	18.2	10.2	10.7	10.0	6.5	0.2	5.00	0.75
WB	10.7	16.8	11.7	14.3	10.8	2.3	0.7	0.75	0.75
LK	12.6	22.5	12.9	14.0	12.6	7.6	0.2	5.25	0.75
LB	16.3	27.6	17.6	16.5	16.3	10.3	1.2	5.25	0.75

Note: Each observation is a (market, route, five-minute interval, day). Southbound peak observations are from 5 AM to 11 AM; off-peak observations are from 11 AM to 7 PM. Markets are ordered from shortest to longest. The three shortest southbound markets do not have a feasible HOT route. The Totem Lake to Bellevue market appears in two rows, labeled TB 1 and TB 2, because it has two HOT routes. The length in miles is equal to the travel time in minutes when traveling at the 60 mph speed limit.

Figure A.2: Travel time and price variation in two markets

(a) Travel time: short market

(b) Travel time: long market

(c) Price: short market

(d) Price: long market
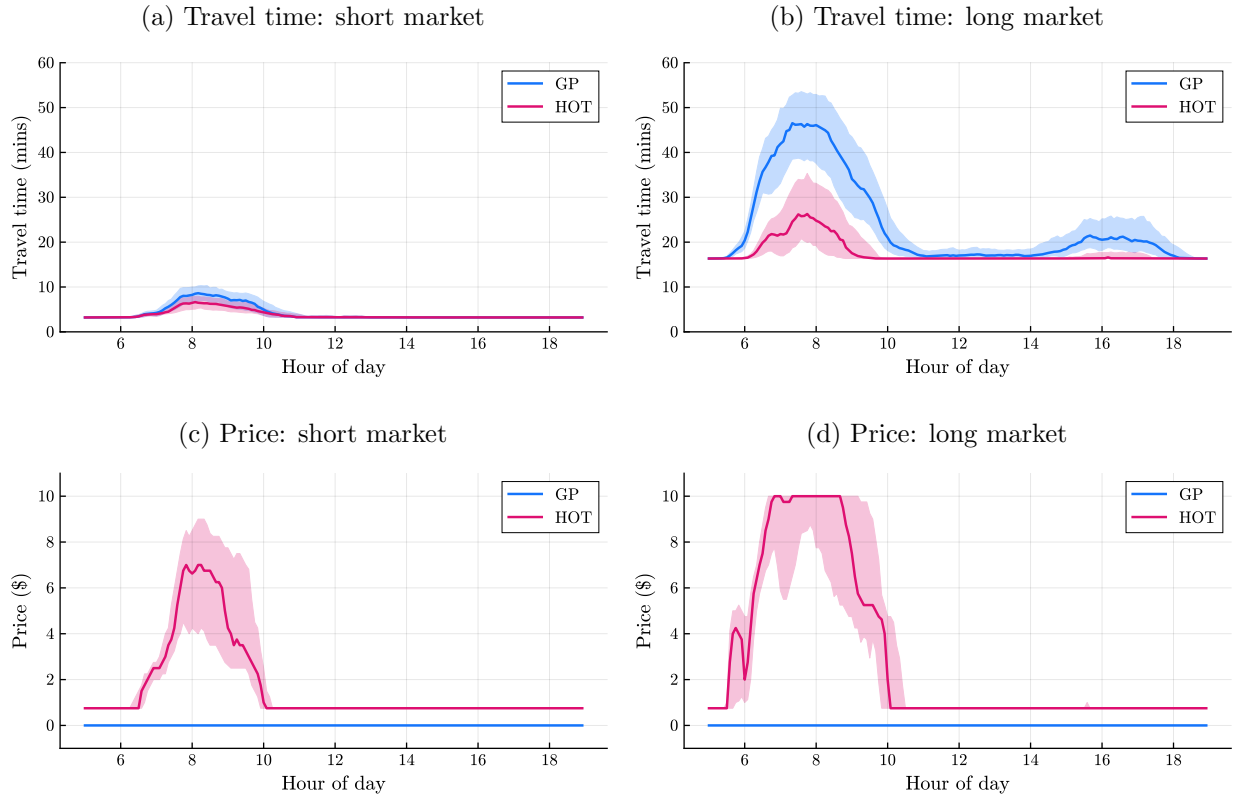
Note: Figures show variation in travel time and price by route and time of day. In each five-minute interval, the thick line indicates the across-day median and the shaded area is between the 25th and 75th percentiles. Each underlying observation is a (market, route, five-minute interval, day) from 5 AM to 7 PM (tolled hours) in 2019. The short market is Totem Lake to Kirkland, the shortest southbound market with a feasible HOT route. The long market is Lynnwood to Bellevue, the longest southbound market.

45

A.2 Descriptive evidence

Figure A.3: Long-run changes in speed by road segment
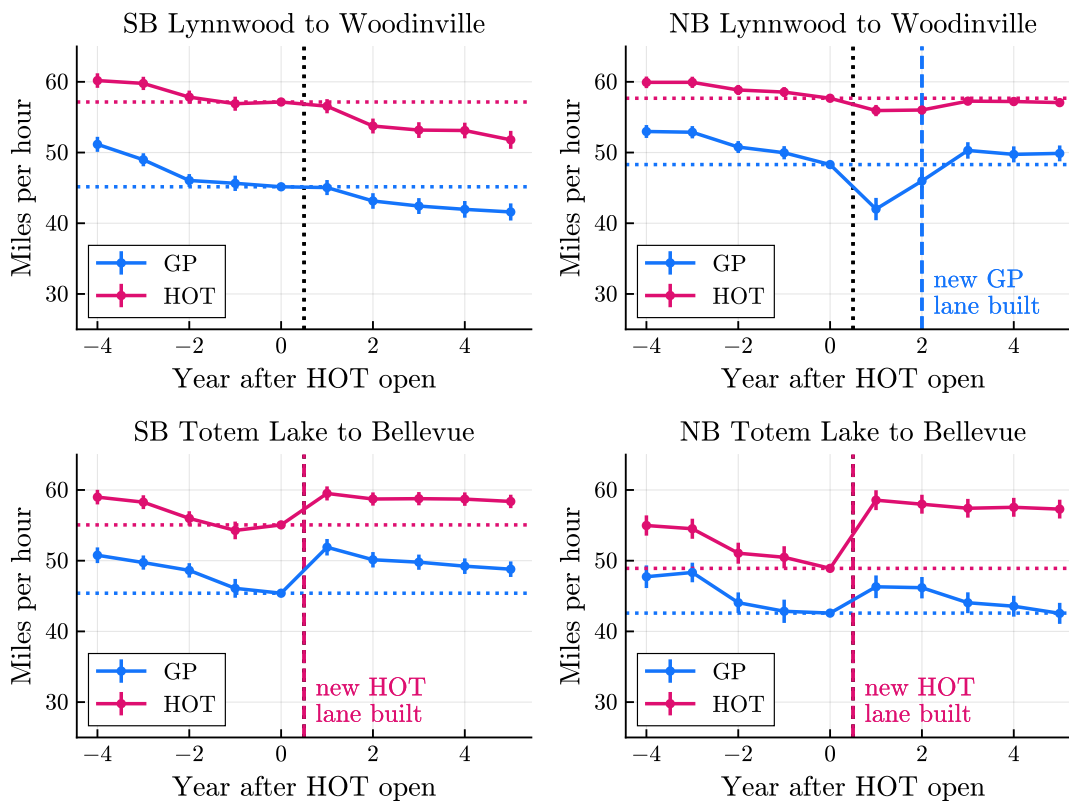
Note: These figures report changes in average speed by year relative to HOT opening, (GP or HOV/HOT) route, and road segment. Each point is an estimated coefficient on a route-year interaction in equation (1), with the level normalized to the GP speed in the last year before the HOT lanes opened. The error bars show 95 percent confidence intervals. The pink dashed lines indicate the September 2015 construction of an additional HOT lane on the southern half of the corridor. The blue dashed lines indicate the April 2017 opening of a GP peak-use shoulder lane on a northbound segment in the northern part of the corridor. The sample is peak hours, southbound 5–11 AM and northbound 1–7 PM, on weekdays from 2011–2019.

Figure A.4: Long-run changes in throughput by road segment
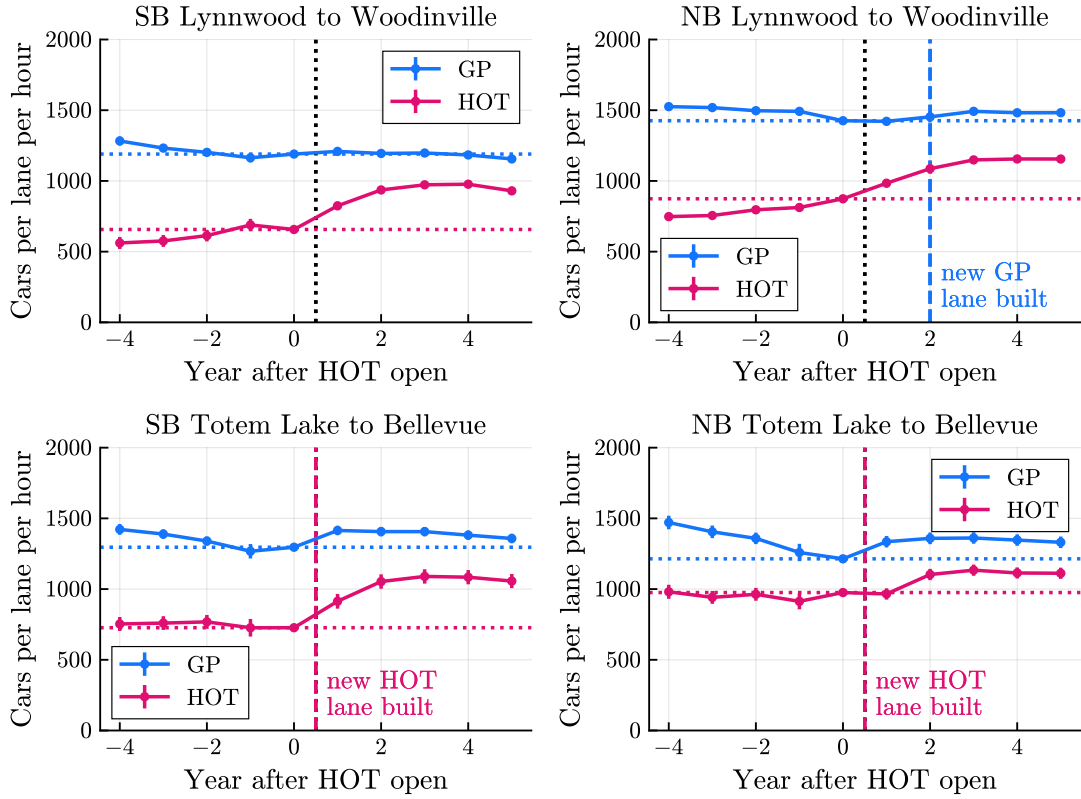
Note: These figures report changes in average throughput by year relative to HOT opening, (GP or HOV/HOT) route, and road segment. Each point is an estimated coefficient on a route-year interaction in equation (1), with the level normalized to the GP throughput in the last year before the HOT lanes opened. The error bars show 95 percent confidence intervals. The pink dashed lines indicate the September 2015 construction of an additional HOT lane on the southern half of the corridor. The blue dashed lines indicate the April 2017 opening of a GP peak-use shoulder lane on a northbound segment in the northern part of the corridor. The sample is peak hours, southbound 5–11 AM and northbound 1–7 PM, on weekdays from 2011–2019.

Figure A.5: Driver markets by car price quartile
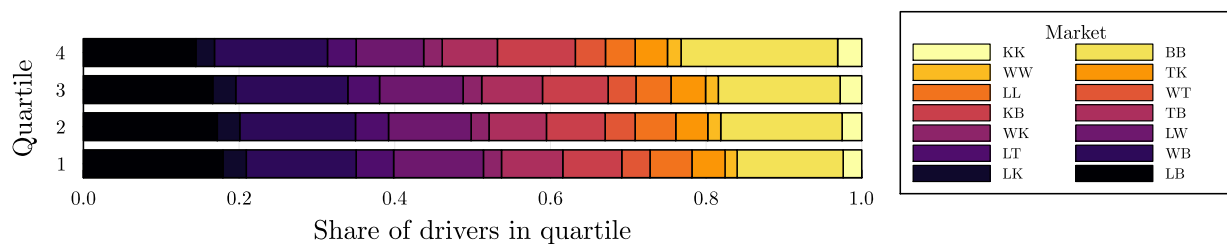
Note: Each underlying observation is a simulated driver in the unconditional sample of potential I-405 drivers. Markets are ordered from shortest (lightest color) to longest (darkest color).

Table A.2: HOT trip attributes vs. driver income proxies

	(1)	(2)	(3)
	Time saved	Price paid	Log price per min saved
Controls: none			
Tract income	−0.0129	−0.0088	−0.0004
	(0.0001)	(0.0001)	(0.0000)
Car price	−0.0095	−0.0028	0.0008
	(0.0002)	(0.0001)	(0.0000)
Controls: market			
Tract income	0.0007	0.0015	0.0000
	(0.0001)	(0.0001)	(0.0000)
Car price	0.0006	0.0018	0.0002
	(0.0002)	(0.0001)	(0.0000)
Controls: market × hour			
Tract income	−0.0018	0.0001	0.0001
	(0.0001)	(0.0000)	(0.0000)
Car price	−0.0009	0.0010	0.0003
	(0.0001)	(0.0001)	(0.0000)

Note: Each underlying observation is a paid peak-hour toll transaction. Southbound peak hours are 5 AM to 11 AM; northbound peak hours are from 1 PM to 7 PM. Time saved is in minutes, price paid is in dollars, and the log of price per minute saved is in base 10.

Figure A.6: Travel times and prices by market, 7–8 AM

(a) GP travel time
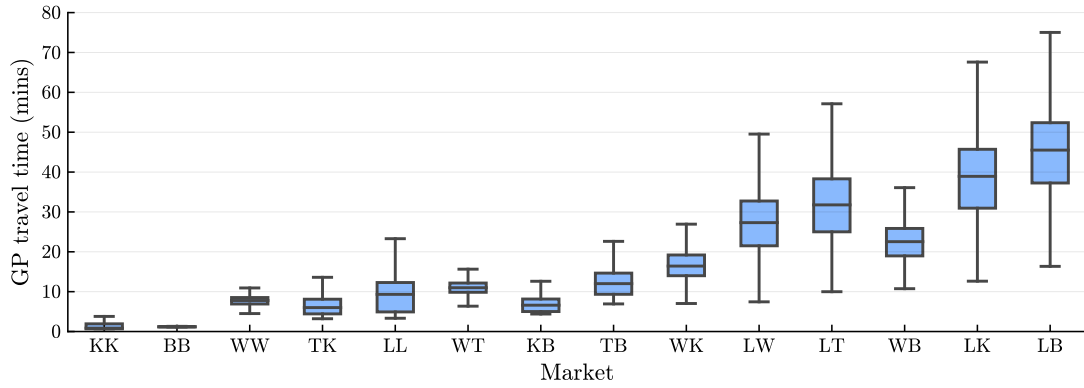
(b) HOT travel time
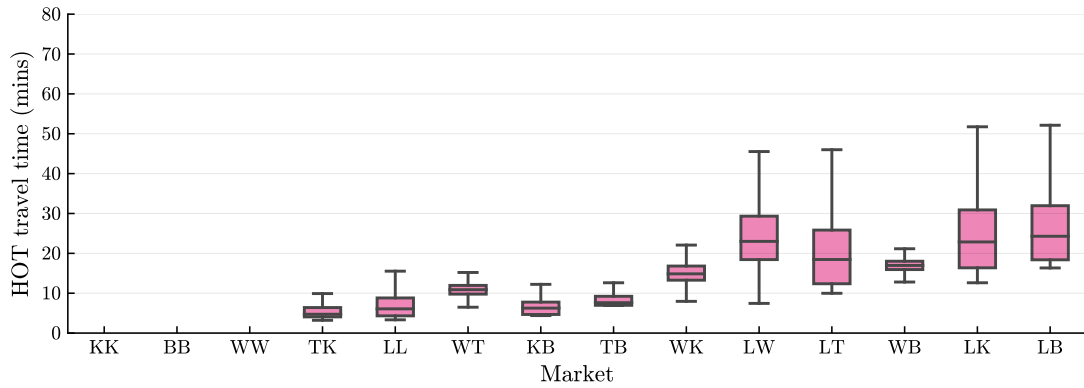
(c) HOT price
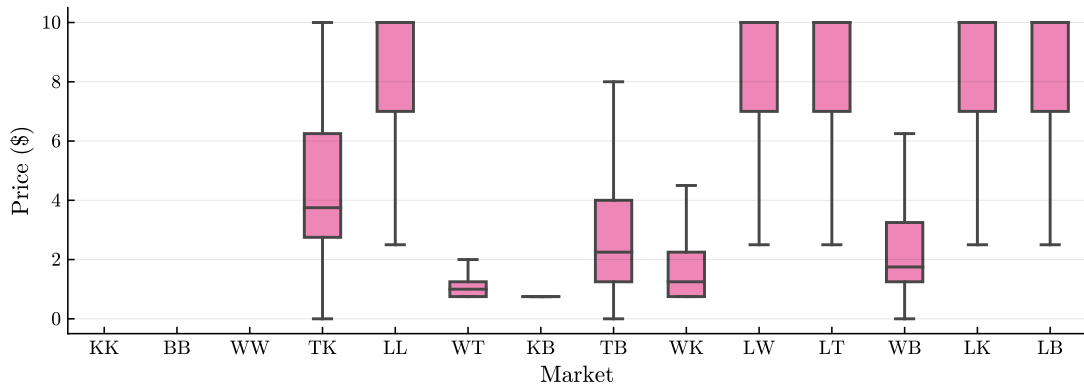
Note: Figures show the distributions of GP travel times (top panel), HOT travel times (middle panel), and HOT prices (bottom panel) from 7–8 AM in each southbound market. Boxes indicate the 25th percentile, median, and 75th percentile. Lower whiskers extend to the lowest observed data point that is within a distance of 1.5 times the interquartile range (IQR) from the 25th percentile. Likewise, upper whiskers extend to the highest observed data point within 1.5 times the IQR from the 75th percentile. Markets are ordered from shortest to longest. The first three markets have no feasible HOT route.

49

A.3 Model

Figure A.7: Example trajectories in spacetime
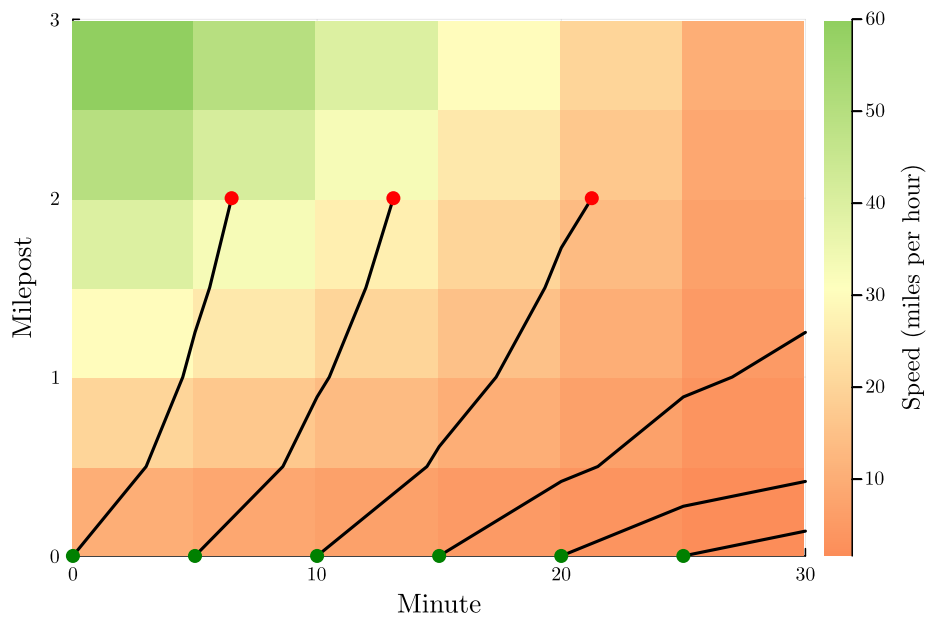
Note: This figure depicts example trajectories for a driver traveling from milepost 0 to milepost 2 at different starting times. Each trajectory begins at a green dot and ends at a red dot. In each shaded rectangle, the slope of the driver's trajectory equals the speed at that discretized location and time.

A.4 Estimation

Figure A.8: Realized HOT time savings by distance to price rounding threshold
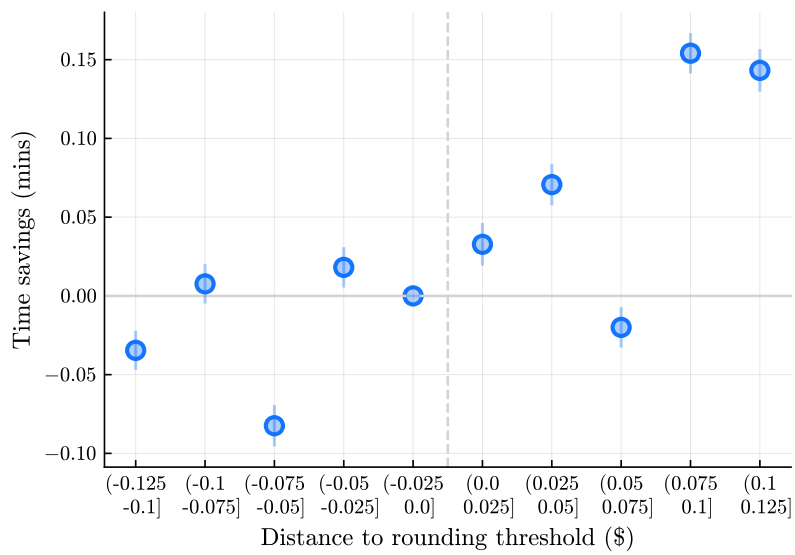
Note: This figures plots coefficients from a regression of HOT time savings on indicators for bins of the unrounded price's distance to the rounding threshold, estimated with threshold and (trip definition, hour) fixed effects. Observations are at the (trip definition, five-minute interval, day) level. Data are subset to southbound 5–11 AM and northbound 1–7 PM.

Figure A.9: Price rounding discontinuity: trip definitions with travel time signs

(a) HOT quantity

(b) Sign-reported HOT time savings
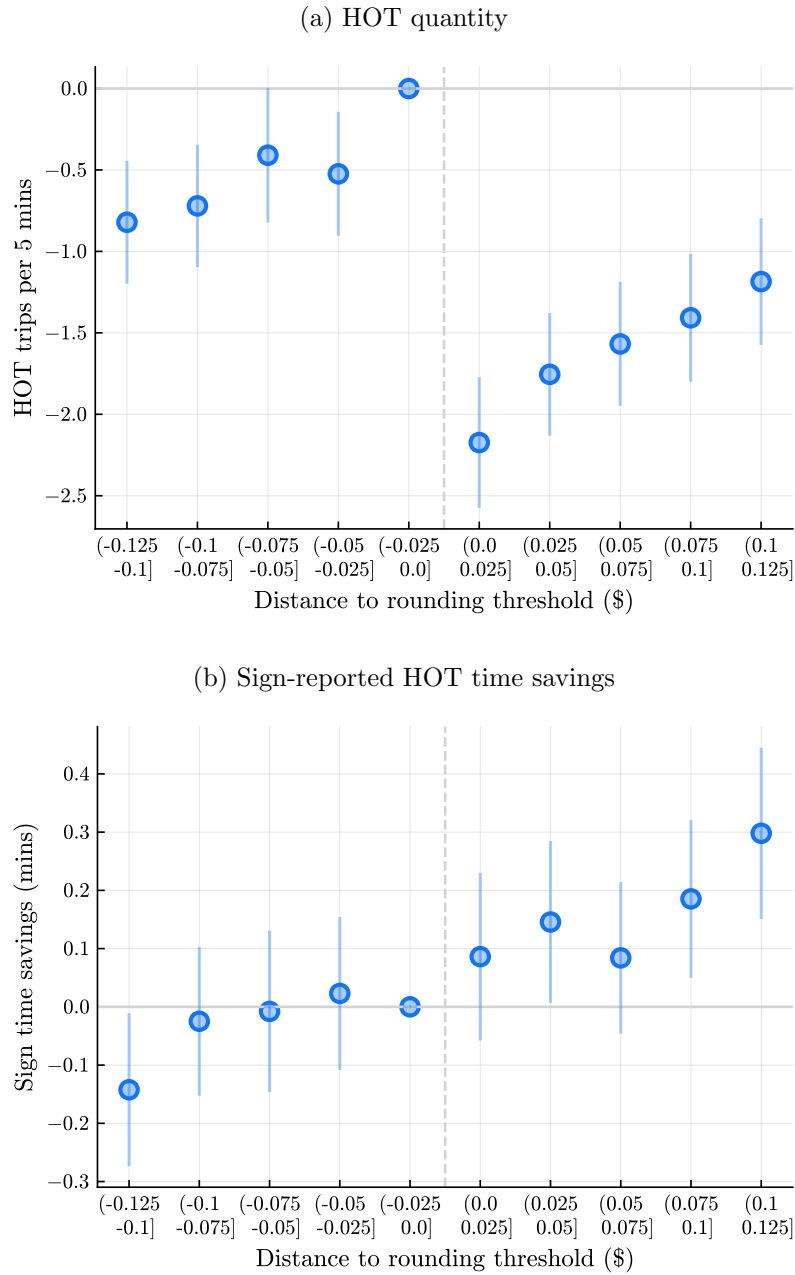
Note: These figures plot coefficients from regressions of HOT quantity (top panel) and sign-reported HOT time savings (bottom panel) on indicators for bins of the unrounded price's distance to the rounding threshold, estimated with threshold and (trip definition, hour) fixed effects, estimated only on trip definitions with travel time signs. Observations are at the (trip definition, five-minute interval, day) level. Data are subset to southbound 5–11 AM and northbound 1–7 PM.

Figure A.10: Travel times and prices with and without precipitation
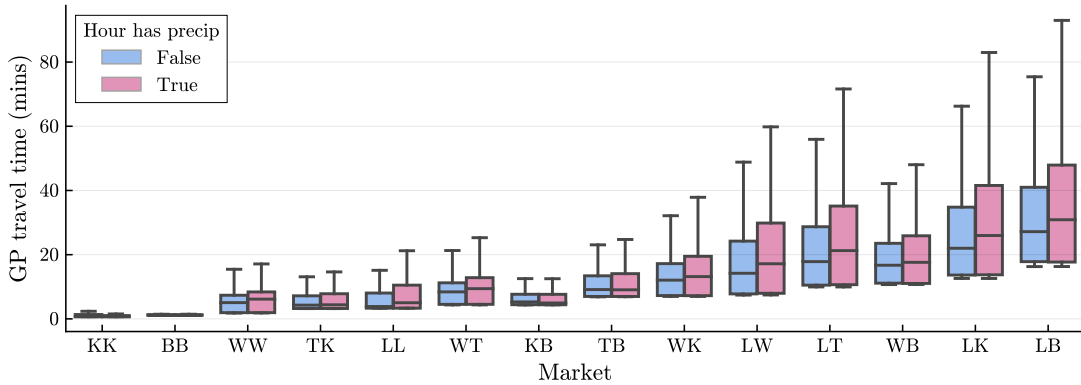
(a) GP travel time
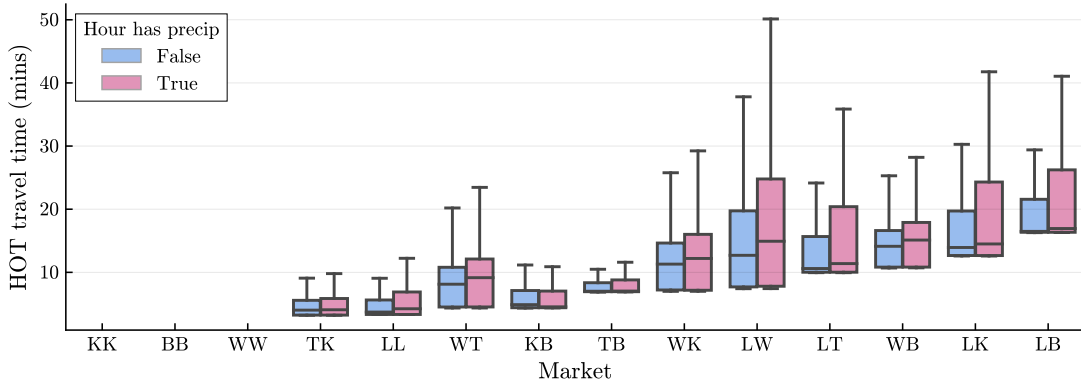
(b) HOT travel time
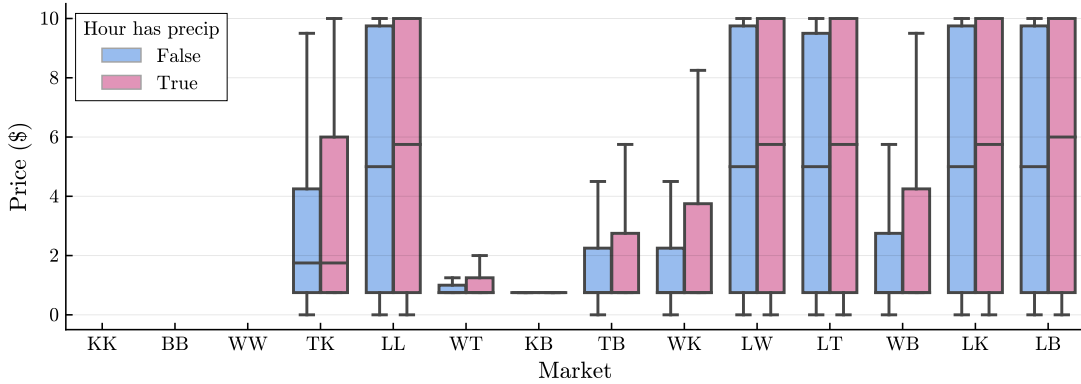
(c) Price

Note: These figure shows the distributions of GP travel times (top panel), HOT travel times (middle panel) and prices (bottom panel) during the morning peak (5–11 AM) in each southbound market, comparing hours with zero precipitation to hours with positive precipitation. Boxes indicate the 25th percentile, median, and 75th percentile. Lower whiskers extend to the lowest observed data point that is within a distance of 1.5 times the interquartile range (IQR) from the 25th percentile. Likewise, upper whiskers extend to the highest observed data point within 1.5 times the IQR from the 75th percentile. Markets are ordered from shortest to longest. The first three markets have no feasible HOT route.

Table A.3: Effects of GP and HOT crashes

	(1) Time savings	(2) Price	(3) HOT quantity
GP crash	1.137 (0.017)	0.723 (0.010)	1.048 (0.038)
HOT crash	0.854 (0.037)	1.207 (0.020)	0.586 (0.057)

Note: This table reports coefficient estimates from equation (7), which includes (trip definition, hour) fixed effects. Price is denominated in dollars and time savings in minutes. Heteroskedasticity-robust standard errors are in parentheses.

Figure A.11: Five-minute departure time shares with and without precipitation

(a) Difference

(b) Levels
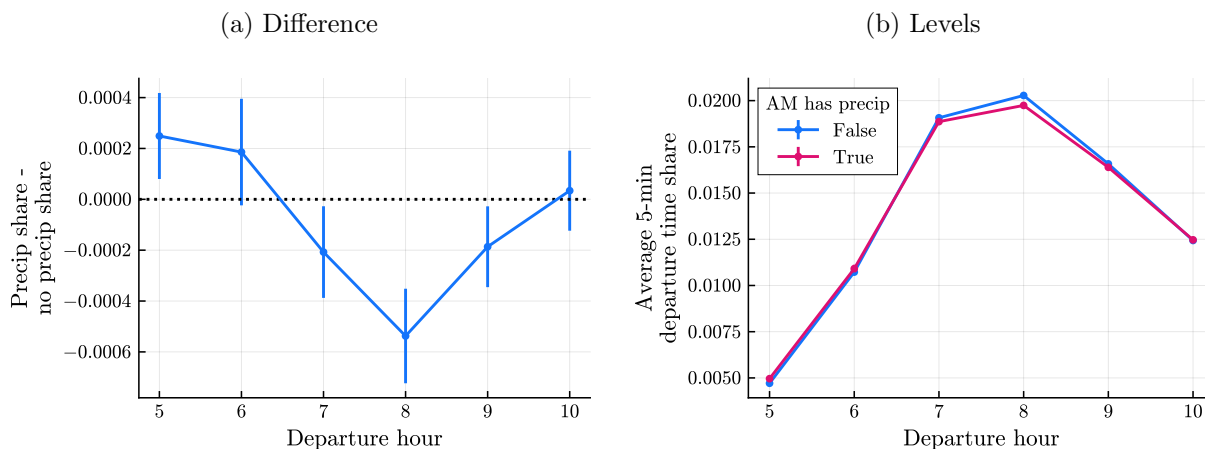
Note: These figures plot coefficients from a regression of departure time shares on hour indicators and hour × AM precipitation interactions, estimated including market origin fixed effects. Each observation is a (market origin, 5-minute interval, date).

Table A.4: Demand parameter estimates

(a) Stage 2 coefficients

	Mean ($\overline{\alpha}$)	Income heterogeneity (μ^α)		Covariance (Σ^α) of unobserved heterogeneity	
		Tract income	Car price	Price	Travel time
Price	-1.112	-0.0014	0.0084	0.0026	0.0179
Travel time	-22.673	-0.0451	0.2890	0.0179	0.6088
Time early	-3.920	–	–	–	–
Time late	-4.817	–	–	–	–

(b) Stage 1 and stage 2 intercepts

	Inside good intercepts (β^0)	HOT intercepts (α^0) by market destination				
		L	W	T	K	B
Lynnwood	8.402	-0.065	-0.950	0.138	2.154	1.652
Woodinville	5.019	–	–	-1.977	1.557	0.716
Totem Lake 1	3.812	–	–	–	-1.334	-0.481
Totem Lake 2	3.812	–	–	–	-1.334	-0.031
Kirkland	-0.815	–	–	–	–	-0.350
Bellevue	-0.464	–	–	–	–	–

(c) Ideal arrival time distribution

Mean ($\overline{\eta}$, not estimated)	8:30 AM
Standard deviation (σ^η)	1.016 hours

Note: These tables report demand parameter estimates. In panel a, the time variables—travel time, time early, and time late—are measured in hours. Tract income and car price are expressed in deviations from the population mean, in thousands of dollars. In panel b, the market origins and destinations are ordered from northernmost to southernmost. Totem Lake appears twice because the Totem Lake to Bellevue market has two HOT routes.

Figure A.12: Value of travel time vs. income proxies
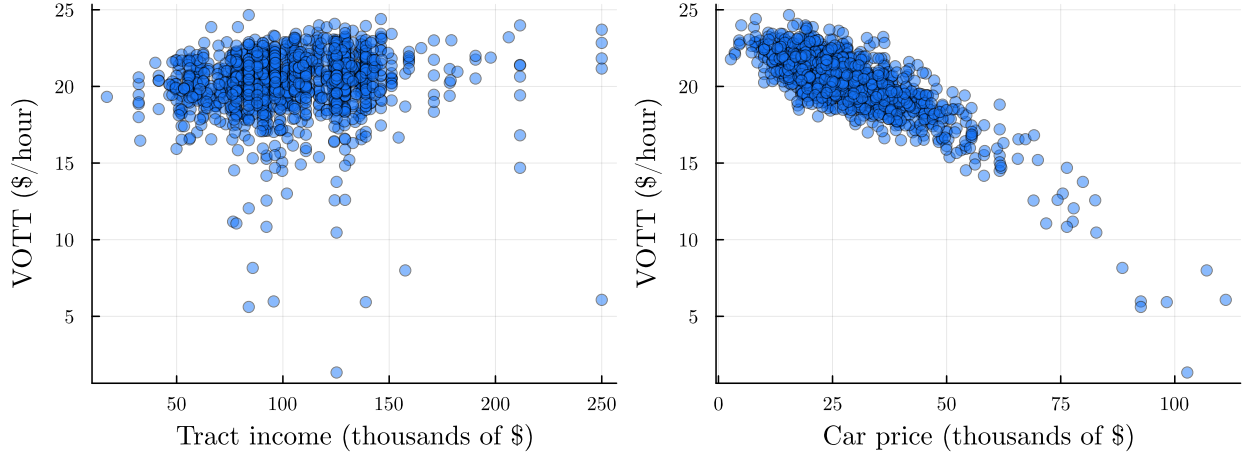
Note: Figures show how drivers' values of travel time vary with their tract incomes and car prices. Each point is a simulated driver from the unconditional distribution of driver characteristics.

Table A.5: Speed-density relationship parameter estimates

Param	Description	Estimate
$\underline{v}$	Jam speed	7.567
		(0.017)
$\overline{v}$	Freeflow speed	63.932
		(0.006)
δ_1	Scale	2.495
		(0.012)
δ_2	Asymmetry	0.100
		(0.001)
ρ^*	Critical density	26.194
		(0.015)

Note: This table reports estimates of the parameters of the asymmetric logistic speed-density relationship (9), estimated via nonlinear least squares. Heteroskedasticity-robust standard errors are in parentheses.

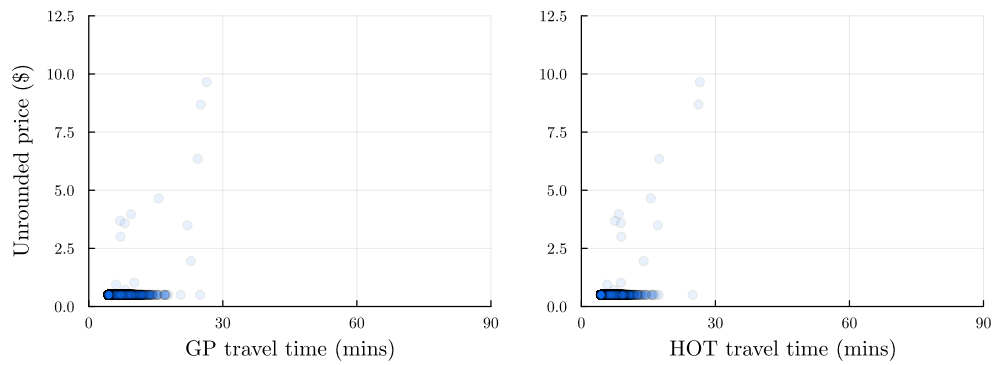Figure A.13: Travel time vs. unrounded price

(a) Lynnwood to Bellevue
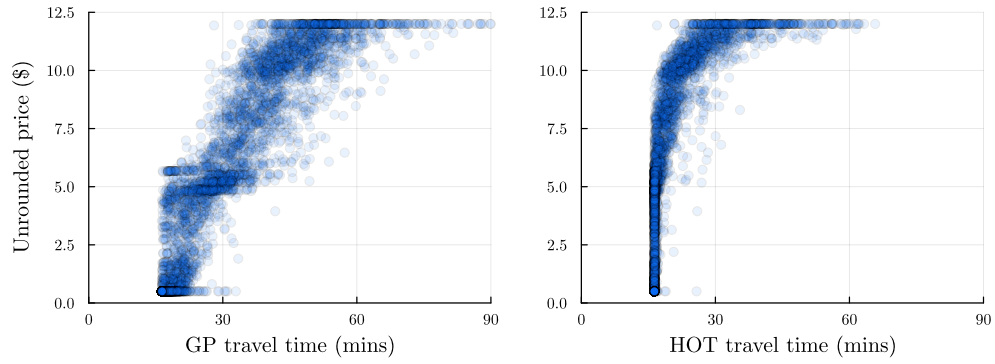
(b) Totem Lake to Bellevue
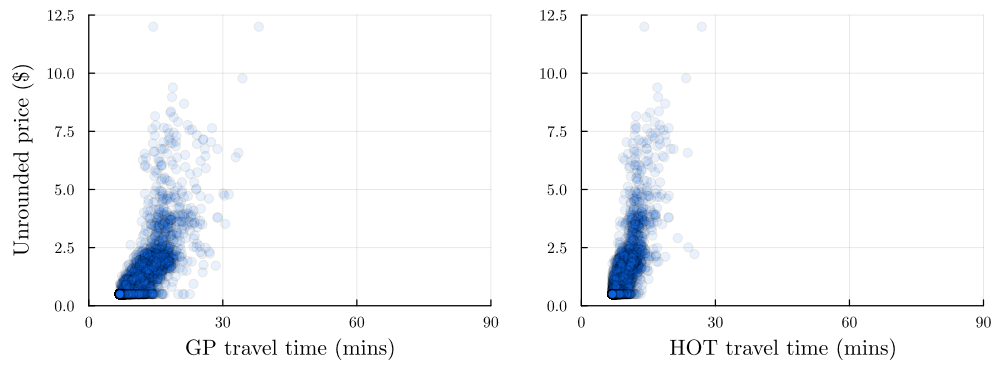
(c) Kirkland to Bellevue

Note: Each figure shows a random sample of 5000 observations, where each observation is a (market, five-minute interval, date) from 5–11 AM in 2019. The markets are ordered from longest to shortest.

A.5   Welfare analysis

Table A.6: No toll vs. status-quo tolling differences by car price quartile

|	Q1	Q2	Q3	Q4
Stage 1 shares				
Outside option	+0.033	+0.037	+0.041	+0.035
5–6 AM	−0.001	−0.002	−0.001	−0.001
6–7 AM	−0.007	−0.005	−0.006	−0.003
7–8 AM	−0.008	−0.012	−0.012	−0.006
8–9 AM	−0.009	−0.012	−0.013	−0.012
9–10 AM	−0.003	−0.004	−0.005	−0.008
10–11 AM	−0.005	−0.003	−0.003	−0.004
Stage 2 shares				
GP	−0.282	−0.284	−0.265	−0.285
HOT	+0.282	+0.284	+0.265	+0.285
Average I-405 outcomes				
Price paid ($)	+0.769	+0.789	+0.704	+0.801
Travel time (mins)	−1.433	−1.371	−1.240	−1.467
Time early (mins)	−2.229	−2.403	−2.100	−2.224
Time late (mins)	−0.761	−1.031	−0.762	−0.873
Welfare				
Driver surplus ($K/day)	−5.840	−7.435	−7.715	−8.072
Toll revenue ($K/day)	+9.202	+9.039	+9.371	+9.979
Net welfare ($K/day)	+2.935	+2.529	+1.702	+1.362

Note: This table reports differences in outcomes between the no-toll counterfactual and status-quo tolling (corresponding to the third column of Table 2) separately by quartiles of car price. The quartiles generate different amounts of toll revenue, but the net welfare values assume uniform redistribution of that toll revenue across all drivers.

Table A.7: No-toll outcomes by tract income quartile

	Q1	Q2	Q3	Q4
Stage 1 shares				
Outside option	0.573	0.413	0.426	0.432
5–6 AM	0.069	0.041	0.035	0.040
6–7 AM	0.085	0.065	0.063	0.069
7–8 AM	0.081	0.135	0.136	0.128
8–9 AM	0.078	0.155	0.138	0.144
9–10 AM	0.063	0.122	0.095	0.113
10–11 AM	0.052	0.068	0.106	0.075
Stage 2 shares				
GP	1.000	1.000	1.000	1.000
HOT	0.000	0.000	0.000	0.000
Average I-405 outcomes				
Price paid ($)	0.000	0.000	0.000	0.000
Travel time (mins)	21.753	16.100	11.810	9.816
Time early (mins)	38.691	24.686	19.620	16.795
Time late (mins)	16.164	12.270	11.662	10.987
Welfare				
Driver surplus ($K/day)	57.229	78.944	50.948	49.192
Toll revenue ($K/day)	0.000	0.000	0.000	0.000
Net welfare ($K/day)	57.229	78.944	50.948	49.192

Note: This table reports outcomes in the no-toll counterfactual in levels (corresponding to the first column of Table 2), separately by quartiles of tract income. The quartiles generate different amounts of toll revenue, but the net welfare values assume uniform redistribution of that toll revenue across all drivers.

Figure A.14: Ex ante and option values of tolling by market
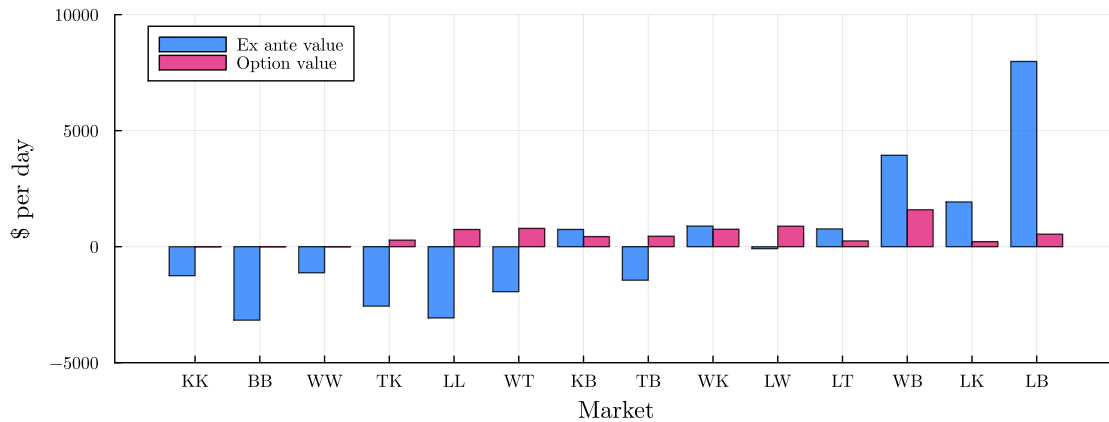
Note: This figure shows differences in aggregate welfare between the no-toll counterfactual and the status quo, decomposed into ex ante and option values of tolling, separately for each southbound market. Markets are ordered from shortest to longest.

B   Data construction

B.1   Market definitions

In order to define our set of markets, we introduce some additional terminology.

An **interchange** is a location on the highway where drivers can travel between I-405 and other roads or highways. Moving from north to south in Figure B.15, the interchanges on the tolled section of I-405 are: I-5, SR 527, NE 195th St, SR 522, NE 160th St, NE 128th St, NE 124th St, NE 116th St, NE 85th Pl, NE 70th Pl, SR 520, NE 8th St, NE 6th St, and NE 4th St. Each interchange may include multiple on- and off-ramps. For example, the SR 520 interchange has on- and off-ramps to and from both eastbound and westbound SR 520.

We call a group of interchanges a **town**. We define the following five towns, again moving from south to north:

- Lynnwood: all points north of the I-5 interchange on SR 525, I-5, and SR 527
- Woodinville: NE 195th St, SR 522, and NE 160th St
- Totem Lake: NE 128th St, NE 124th St, and NE 116th St
- Kirkland: NE 85th Pl and NE 70th Pl
- Bellevue: SR 520, NE 8th St, NE 6th St, NE 4th St, and all points south of NE 4th St on the untolled section of I-405

Finally, a **market** combines a direction of travel with an entry town and an exit town. The entry town and exit town may be the same: for example, travel from I-5 to SR 527 occurs in the southbound Lynnwood to Lynnwood market. The exception is travel from Totem Lake to Totem Lake, which is infeasible in both directions. There are therefore fourteen markets in each direction.

Taking the HOT route is not feasible in every market. We say that the HOT route is **feasible** in a given market if there is a way for drivers to enter the highway at an on-ramp, enter the HOT lanes at a legal access point (indicated by a white triangle in Figure B.15), exit them at a legal access point (indicated by a black triangle in the same figure), and finally exit the highway at an off-ramp. The HOT route is feasible in the southbound Lynnwood to Lynnwood market: drivers can enter I-405 at the I-5 interchange, enter the HOT lanes at the topmost white triangle, exit them at the topmost black triangle, and finally exit I-405 at the SR 527 interchange. It is not feasible in the southbound Bellevue to Bellevue market, as there are no legal HOT entry points south of SR 520.

B.2   Approximating density using speed and throughput

While the induction loops do not record densities, we approximate density at each loop in each five-minute interval by dividing throughput by speed. Identity (11) specifies an *instantaneous* relationship between the speed, density, and throughput at each point in space-time. Our density calculation is an approximation because we are using time-averaged measurements of speed and throughput. This approximation is very accurate at higher speeds, when vehicles are nearly uniformly distributed throughout both space and time, and slightly worse at lower speeds (Hall, 2005).

Figure B.15: I-405 Express Lanes map

Figure B.16a shows the underlying speed and throughput data.

To validate our approximation, we compare our estimated densities to observed loop occupancy rates, which are measured directly by the induction loops. The occupancy rate is the fraction of time the loop has a vehicle passing over it. In theory, density equals occupancy multiplied by the average vehicle length. Figure B.16b shows that our estimated density is indeed linear in the observed occupancy rate. The R^2 from regressing observed occupancy on estimated density is 0.975.

Figure B.16: Relationships between loop traffic variables

(a) Throughput vs. speed

(b) Density vs. occupancy
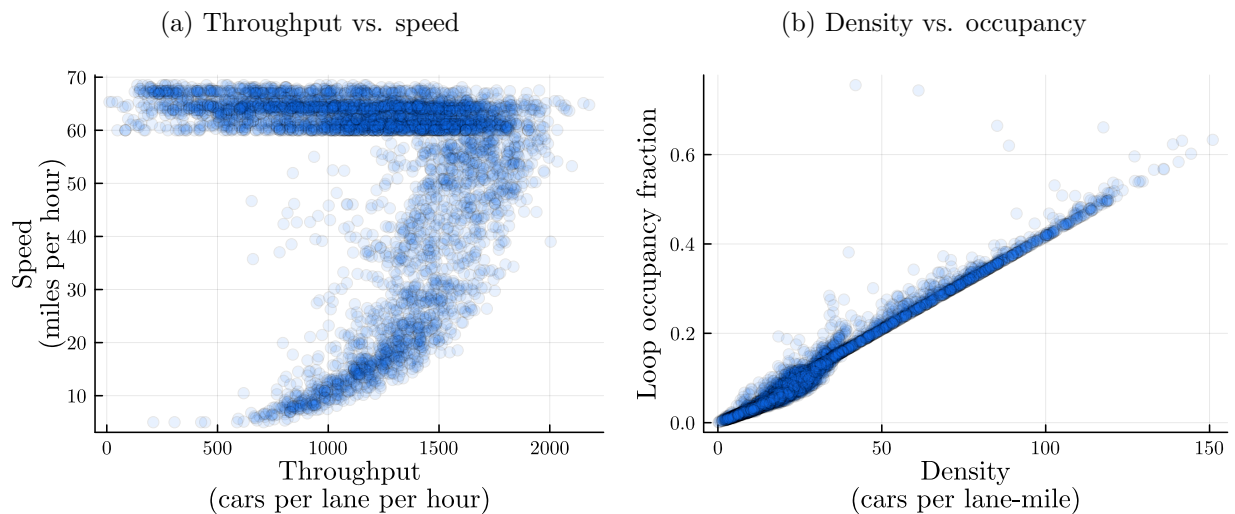
Note: Figures show a random sample of 5000 southbound observations, where each observation is a (loop, five-minute interval, date) from 5–11 AM in 2019. Top-coded speeds have been replaced with estimated freeflow speeds.

B.3 Estimating vehicle MSRPs

We use transaction-level data on vehicle registrations in Washington State, obtained via a public disclosure request from the Washington State Department of Licensing. For each transaction, we observe attributes including the date of registration; the make (e.g., Toyota), model (e.g., Prius), and model year of the vehicle being registered; the home Census tract of the registrant; and the amount of motor vehicle excise tax (MVET) paid. Our sample contains the universe of over 38 million vehicle registrations in the state from January 2017 to December 2022. Vehicle owners are required to register their vehicles when they move to the state and to renew their registrations annually; both types of transactions require MVET payments.

For each Seattle-area registration, we compute the manufacturer-suggested retail price (MSRP) implied by the amount of MVET paid. The MVET is levied on vehicle registrants living within the Sound Transit District. This district covers parts of three Seattle-area counties served by the Central Puget Sound Regional Transit Authority; it contains the I-405 Express Toll Lanes that we study in this paper. The MVET is a fraction of the vehicle's depreciated value, which is in turn

computed from the vehicle's MSRP and a depreciation schedule (Sound Transit, 2023). We use the 13 million registrations with positive MVETs (i.e., registrations occurring in the Sound Transit District) beginning in March 2017, when the excise tax rate increased to 1.1 percent. At this stage, we discard the 0.0025 percent of registrations where the estimated MSRP is more than \$500,000; these observations likely reflect data errors or mistakes made by the registrant. Figure B.17 shows the distributions of estimated MSRPs for four sample makes.

Figure B.17: Estimated MSRPs: vehicles with model year 2015 registered in 2019
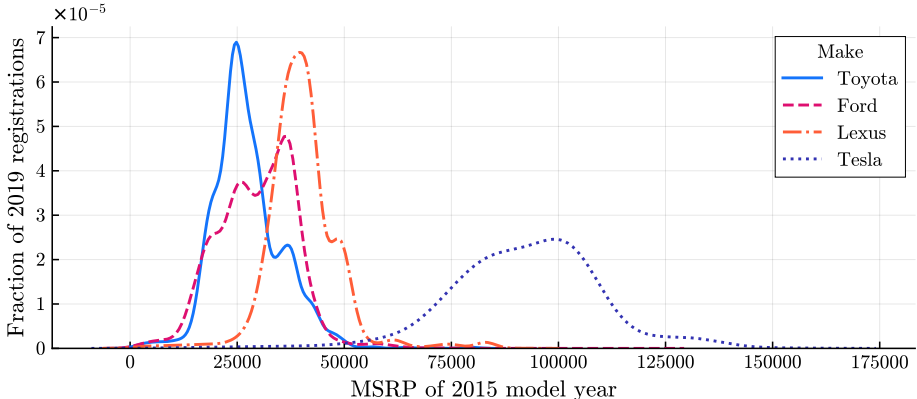
Finally, we match each I-405 HOT driver's car to an MSRP estimate, which we use as our measure of the car price. If possible, we match the car to the median MSRP estimated for vehicles with the same make, model, and year. If not, we match it to the median MSRP among registered vehicles with the same make and model. If still not, we use the median MSRP among registered vehicles with the same make. Among HOT drivers with vehicle information, 73 percent are matched to the make, model, and year; 7 percent are matched to the make and model only; 2 percent are matched to the make only; and the remaining 1 percent are unmatched.

B.4 Market sizes and characteristics

Let $r \to s$ denote an ordered pair of Census tracts. We define tract pair $r \to s$ as belonging to market m if one of the top three driving routes from tract r to tract s (according to the OpenStreetMaps router) involves taking I-405 in market m. Each tract pair belongs to at most one market.

First, we compute the size of each market m on each date t using the Replica GPS data. Slightly abusing notation, let $q_{r \to s, H, t}$ denote the historical travel flows from tract r to tract s in hour H on date t. Replica estimates these quantities based on trips taken by a sample of GPS devices. We compute q_{mt}, the size of market m on date t, by summing over tract pairs in the market and over hours $H \in \{5, \ldots, 10\}$ in the morning peak:

$$q_{mt} = \sum_{(r \to s) \in m} \sum_{H} q_{r \to s, H, t}$$

Figure B.18 illustrates this procedure for m = Lynnwood to Bellevue.

Figure B.18: Example market size construction: Lynnwood to Bellevue
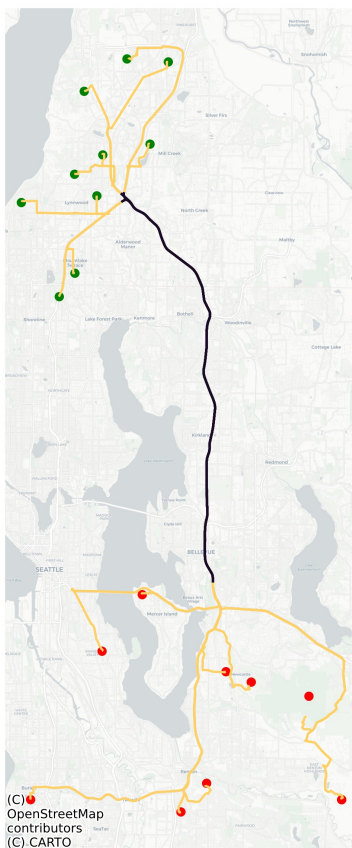
Note: Each green dot is the centroid of an origin Census tract. Each red dot is the centroid of a destination Census tract. The figure shows a sample of the tract pairs for which taking I-405 southbound from Lynnwood to Bellevue (the full length of the tolled section) is one of the top three suggested routes. The black path indicates the part of the route on I-405.

Next, we construct the joint distribution of tract income X and car price Y in each market m.[45] Each origin tract r is associated with a unique tract income value x_r and a distribution $G_r^Y(y)$ of car prices, from registrations in that tract. Let $G_r^{X,Y}(x,y)$ denote the joint distribution of tract income and car price in tract r. Let ω_{rm} denote the share of total market m travel that originates in tract r:

$$\omega_{rm} = \frac{\sum_s \sum_H \sum_t q_{r \to s, H, t}}{\sum_t q_{mt}}$$

The joint distribution $F_m^{X,Y}(x,y)$ of tract income and car price in market m is the weighted average

---

[45] We can't get this distribution directly from the toll transaction data because that sample includes only drivers who have taken the toll lanes at least once.

of that joint distribution in each origin tract:

$$F_m^{X,Y}(x,y) = \sum_r \omega_{rm} G_r^{X,Y}(x,y)$$

B.5   On-ramp metering

As discussed in Section 3.2, each on-ramp contains both metered (general-purpose) and unmetered (carpool-only) lanes, and the metered throughputs reflect rationed demand for I-405 departure times.[46] True demand for 8 AM departures may be higher than observed 8 AM throughput because non-carpooling drivers must queue in order to exit the off-ramp and enter the highway. This queueing represents a loss of information from the researcher's perspective: a single queue exit rate profile can be rationalized by potentially many queue entry rate profiles.

Figure B.19 shows intraday variation in the distributions of speeds and throughputs in the metered and unmetered lanes of an example southbound on-ramp. In the unmetered lane, speed is always constant at (the topcoded value of) 60 miles per hour. Throughput varies smoothly throughout the day, peaking in the morning between 7–8 AM and largely paralleling the profile of intraday travel time variation in Figure A.2. In contrast, in the metered lane, speed drops discontinuously to 10 miles per hour from 6–10 AM. Throughput increases from 5–6 AM, but is flat from 6–10 AM while metering is in place. Throughput then remains elevated after 10 AM, due to a combination of queue emptying and true demand for post-10 AM departure times.

Our approach is to reallocate the metered quantities so that they match the profile of unmetered quantities within each day and ramp. To formalize this, we first define some notation which will be used in this appendix only. Let $\widetilde{\mathcal{H}}$ denote the set of five-minute intervals during tolled hours between 5 AM and 7 PM. Let $\widetilde{q}_{jhrt}$ denote the raw throughput in *lane type* (unmetered or metered) $j \in \{0,1\}$ at time h on ramp r. We use the raw unmetered quantities q_{0hrt} and impute the metered quantities q_{1hrt} as follows:

$$q_{0hrt} = \widetilde{q}_{0hrt}$$

$$q_{1hrt} = \underbrace{\frac{\widetilde{q}_{0hrt}}{\sum_{h' \in \widetilde{\mathcal{H}}} \widetilde{q}_{0h'rt}}}_{\substack{\text{share of unmetered} \\ \text{departures at time } h \\ \text{on day } t}} \times \underbrace{\sum_{h' \in \widetilde{\mathcal{H}}} \widetilde{q}_{1h'rt}}_{\substack{\text{total metered} \\ \text{departures} \\ \text{on day } t}}$$

Figure B.20 shows the distributions of raw and imputed quantities in each market origin (summing across lane types and on-ramps).

This approach amounts to assuming that single-occupancy drivers demand the same distribution

---

[46]The WSDOT website writes of the design and goals of ramp metering: "Ramp meters are a specific type of traffic signal used to control how quickly vehicles enter traffic flow on a freeway, and are a freeway operation strategy designed to reduce collisions and decrease travel times. Ramp meters function by controlling the rate (metering) at which vehicles enter the freeway... Ramp meters typically operate during peak congestion times: 6 AM to 9 AM, and 3 PM to 7 PM. Meters may still be operated outside these hours, as their operation depends on freeway traffic speeds and volumes, and not on time of day" (Washington State Department of Transportation, 2023).

Figure B.19: Ramp metering example: southbound SR 522 on-ramp

(a) Speed
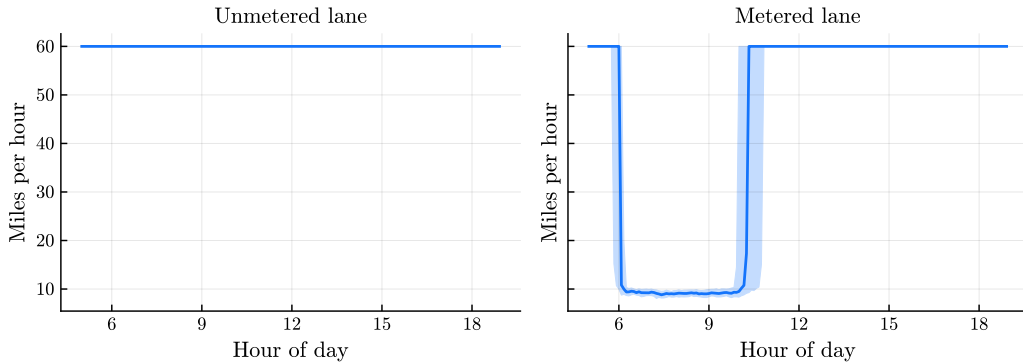
(b) Throughput
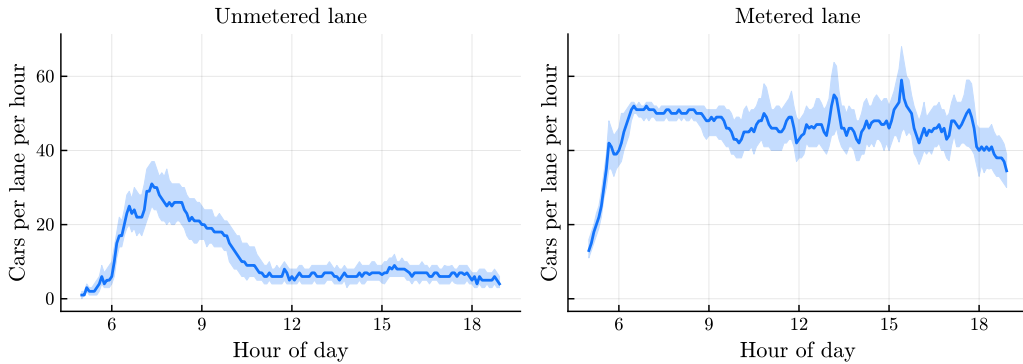
Note: Figures show variation in speed and throughput by (unmetered or metered) lane and time of day. In each five-minute interval, the thick line indicates the across-day median and the shaded area is between the 25th and 75th percentiles. Each underlying observation is a (loop, five-minute interval, day) from 5 AM to 7 PM (tolled hours) in 2019. This ramp, which takes drivers from eastbound SR 522 to southbound I-405, has an unmetered carpool-only lane and a metered general-purpose lane.

of departure times as carpooling drivers. This assumption is potentially violated if carpooling drivers, who face reduced-price or free travel in the HOT lanes (depending on whether they have two or three-plus people in the car), have greater demand for peak-hour travel than single-occupancy drivers. In this case, we will tend to underestimate the average disutilities of price and travel time in the population.

C   Model appendix

C.1   Utility microfoundation

The microfoundation for the second-stage utility specification (2) comes from Vickrey (1969), which posits that drivers have four different flow values of time. Figure C.21 illustrates these flow values, which depend both on where the driver is—at her origin, on the road, or at her destination—and

Figure B.20: Raw vs. imputed departure time quantities
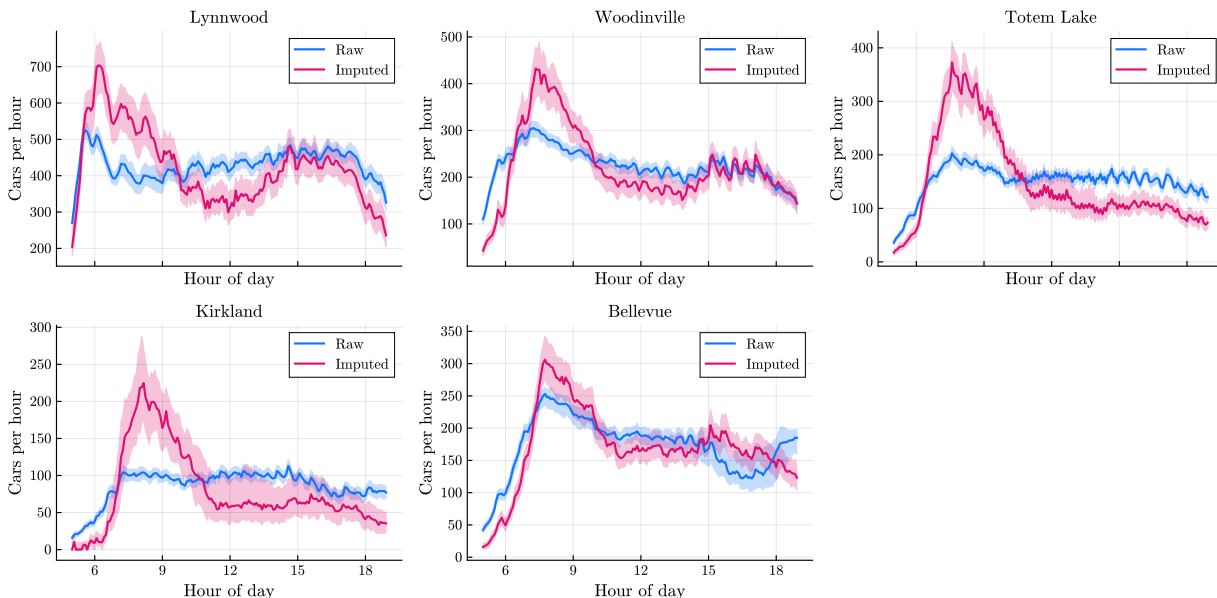
Note: Figures show variation in raw and imputed departure time quantities in each southbound market origin. In each five-minute interval, the thick line indicates the across-day median and the shaded area is between the 25th and 75th percentiles. Each underlying observation is a (market origin, five-minute interval, day) from 5 AM to 7 PM (tolled hours) in 2019.

on the time of day. Regardless of the time of day, driver i derives value u_i^{orig} per hour spent at her origin and u_i^{road} per hour spent on the road. However, her value per hour spent at her destination depends on whether that time is spent before or after ideal arrival time η_i: she gets $\underline{u}_i^{\text{dest}}$ per hour before η_i and $\overline{u}_i^{\text{dest}}$ per hour after η_i.

The utility formulation in equation (2) follows from setting:

$$\alpha_i^D = u_i^{\text{road}} - u_i^{\text{orig}}$$
$$\alpha_i^E = \underline{u}_i^{\text{dest}} - u_i^{\text{orig}}$$
$$\alpha_i^L = u_i^{\text{orig}} - \overline{u}_i^{\text{dest}}$$
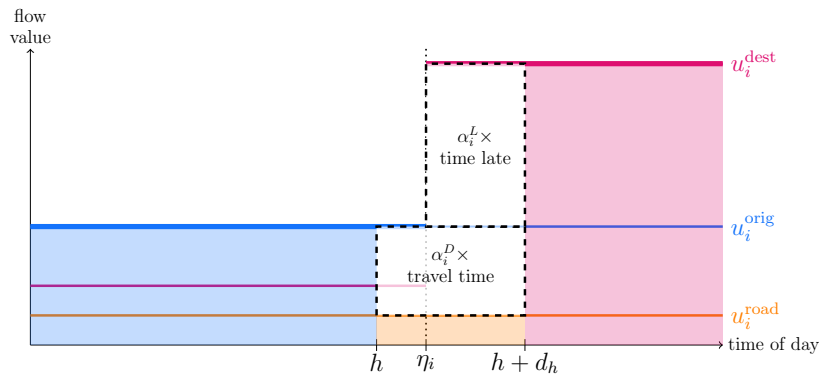
C.2   Speed, density, and throughput

In Section 5.2, we refer to the speed-density relationship as "the" road technology, but in fact, the road technology can be specified as a relationship between any pair of the three variables: speed, density, and throughput. This is because throughput, the number of cars crossing a point in space per unit of time, is identically the product of speed and density, which can be seen in the variables'
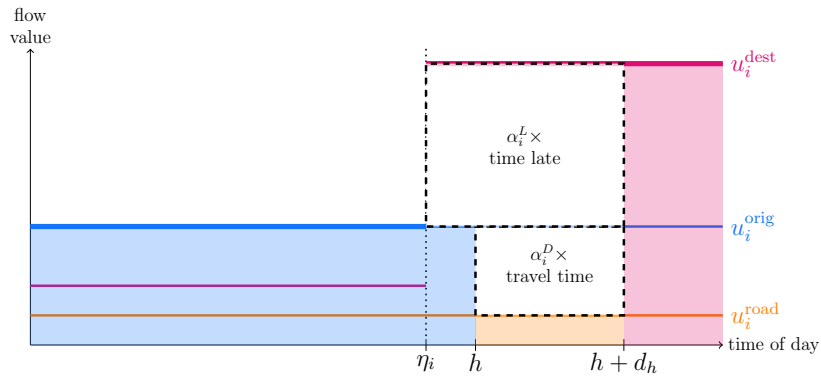
Figure C.21: Utility microfoundation

(a) Case 1: depart early, arrive early
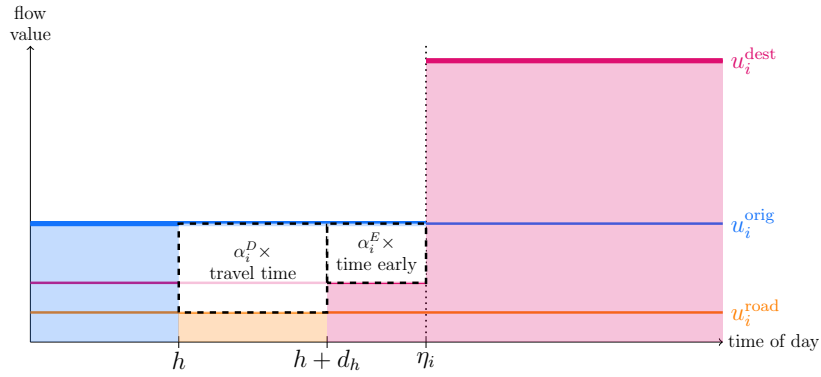
(b) Case 2: depart early, arrive late


(c) Case 3: depart late, arrive late

units of measure:

$$\text{(throughput in cars/lane/hour)} = \text{(speed in miles/hour)} \times \text{(density in cars/lane/mile)} \qquad (11)$$

Thus, given a speed-density relationship, the relationships between the remaining pairs are determined by identity (11).

Figure C.22 illustrates two versions of the road technology which are common in the transportation literature. The speed-density curve (panel a) is monotone, with speed decreasing in density. The speed-throughput curve (panel b) is backward-bending. Start at the top left of the curve, where high speed (and low density, by panel a) is associated with low throughput. As speed decreases, throughput first increases due to increasing density. At some point, however, lower speeds reduce throughput, a phenomenon sometimes referred to as hypercongestion (Hall, 2018; Anderson and Davis, 2020). Both the speed-density and speed-throughput relationships are estimated in the empirical transportation literature, reviewed by Hall (2005).

Figure C.22: Road technology relationships

(a) Speed vs. density

(b) Speed vs. throughput
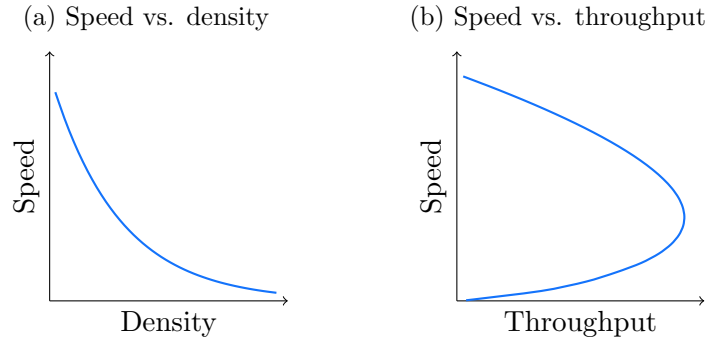
# D   Estimation appendix

D.1   Estimating driver beliefs

In each market m, drivers have common beliefs about the joint distribution of prices and travel times on each day t. Let $(\mathbf{p}_{mt}, \mathbf{d}_{mt}) = (p_{jhmt}, d_{jhmt})_{j \in \{0,1\}, h \in \mathcal{H}}$ denote the $(3\,|\mathcal{H}|)$-dimensional vector of prices and travel times in all routes j and all departure times h. This appendix describes how we parameterize and estimate the joint distribution $G_{mt}(\mathbf{p}_{mt}, \mathbf{d}_{mt})$.

We estimate a truncated joint normal distribution for each market m. The mean and variance-covariance matrix depend on the quarter (i.e., the season), day of week, and presence of absence of precipitation on day t. Estimation proceeds in three steps:

1. Obtain the mean vector by regressing

$$y_{jhmt} \sim \text{departureTime}_h \times \text{dayOfWeek}_t \times \text{quarter}_t \;\; + \;\; \text{departureTime}_h \times \text{precip}_t \qquad (12)$$

Figure D.23: Driver beliefs: mean travel times and prices without precipitation

(a) GP travel time

(b) HOT travel time
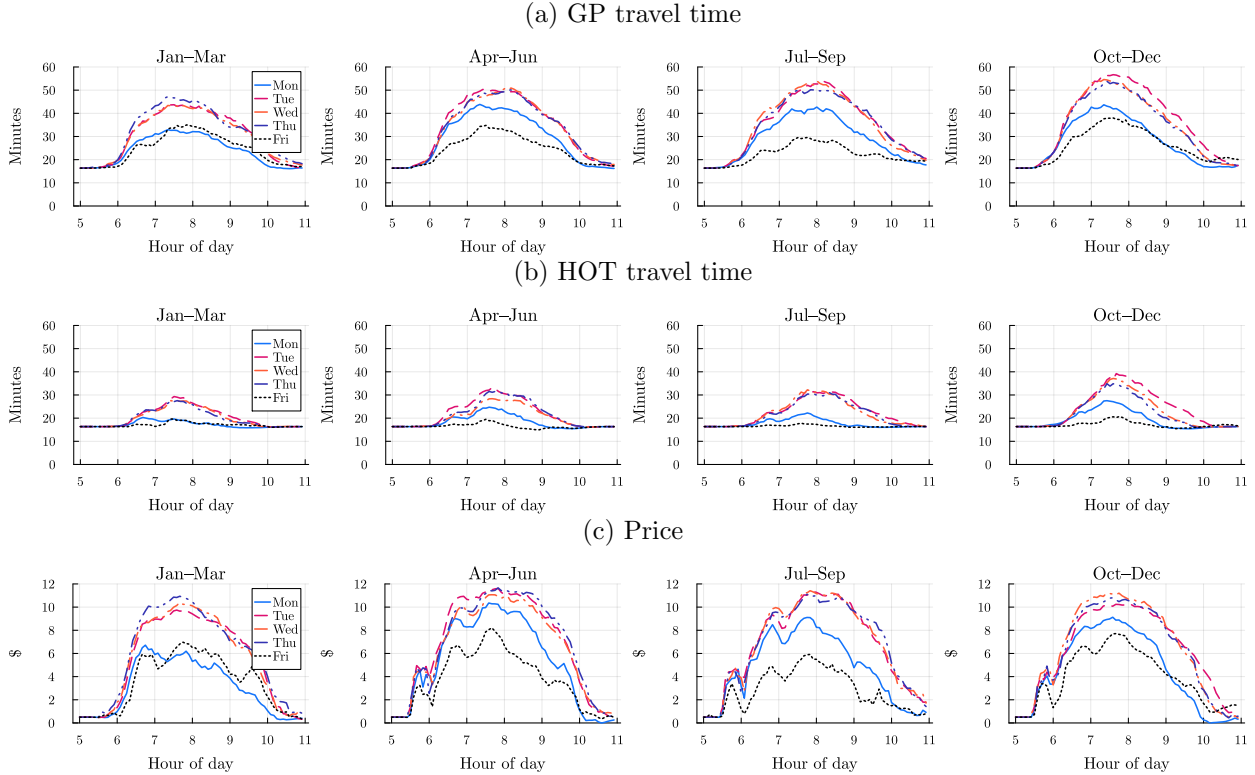
(c) Price

Note: Figure plots estimated coefficients on the (departure time, day of week, quarter) interactions in equation (12).

On the left-hand side, y_{jhmt} is the *unrounded* price or the travel time. On the right-hand side, precip_t is an indicator for positive precipitation at the Everett weather station (about 11 miles north of I-405's northern terminus) between 5–11 AM on day t. Figure D.23 shows the estimated means on days without morning precipitation (i.e., the coefficients on the first right-hand side term).

2. Separately for days with and without precipitation, estimate the joint variance-covariance matrix for $(\mathbf{p}_{mt}, \mathbf{d}_{mt})$ using the sample variance-covariance of the residuals of equation (12). Each variance-covariance matrix has dimension $3\,|\mathcal{H}| \times 3\,|\mathcal{H}|$; Figure D.24 plots the $|\mathcal{H}| \times |\mathcal{H}|$-dimensional submatrices corresponding to the variance-covariance matrices of GP travel times, HOT travel times, and prices.

3. Truncate prices at the \$0.75 floor and \$10 ceiling. Truncate travel times from below at the minimum observed travel time in that market (across all departure times and dates).

D.2 Demand moment conditions

In this appendix, we describe the full set of moment conditions used to jointly estimate the two stages of the demand model. Each set of moment conditions is presented under the heading of

70

Figure D.24: Driver beliefs: variance-covariance matrices of travel times and prices

(a) GP travel time

(b) HOT travel time
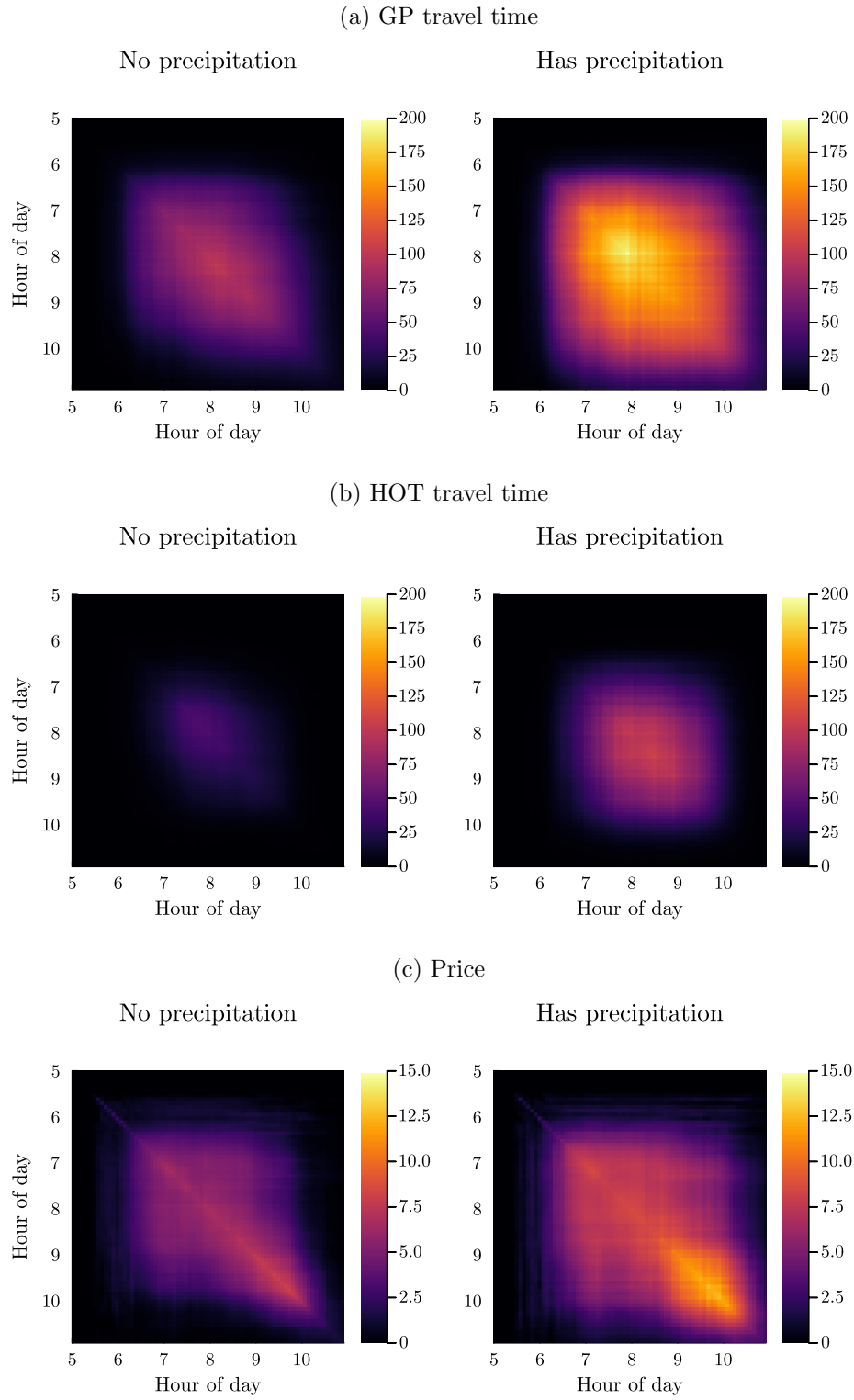
(c) Price

Note: Figures show estimated variance-covariance matrices of the residuals of GP travel times (panel a), HOT travel times (panel b), and prices (panel c) from equation (12).

the demand parameters about which they are most informative. Throughout this appendix, the subscripts underneath an expectation symbol indicate the dimensions over which the expectation is being taken.

D.2.1 Stage 1 moment conditions

Let $\widehat{\xi}_{hat}(\theta)$ denote the demand shocks in market origin a at time h on date t that rationalize the observed departure time shares conditional on the candidate parameters θ. Recall that we can only recover demand shocks at the *market origin*—not market—level because we only observe departure time quantities at the market origin level, from highway on-ramp throughputs.

Furthermore, let $\widetilde{\xi}_{hat}(\theta)$ denote the same demand shock after residualizing with respect to date t fixed effects. These date-demeaned demand shocks affect only substitution across departure times, not substitution toward the non-405 outside option.

Inside good intercepts   The demand shocks have mean zero in each market origin a:

$$\forall a: \qquad \mathbb{E}_{ht}\left[\widehat{\xi}_{hat}(\theta)\right] = 0$$

These moment conditions are informative about the inside good intercepts β_a^0, which control substitution to the non-405 outside option.

Ideal arrival time distribution   After demeaning by date, the demand shocks have mean zero in each hour H:

$$\forall H: \qquad \mathbb{E}_{hmt}\left[\widetilde{\xi}_{hmt}(\theta) \mid \text{hour}(h) = H\right] = 0$$

These moment conditions are informative about the standard deviation σ^η of ideal arrival times. They impose that systematic differences in departure time shares across hours must be due to the distribution of drivers' ideal arrival times, rather than rationalized solely by demand shocks.

Time early and time late coefficients   After demeaning by date, the demand shocks are independent of morning precipitation indicators in each hour H:

$$\forall H: \qquad \mathbb{E}_{hmt}\left[\widetilde{\xi}_{hmt}(\theta) \times \text{precip}_t \mid \text{hour}(h) = H\right] = 0$$

As discussed in Section 6.1.2, precipitation increases the *variance* of prices and travel times. If drivers find it very costly to be late to their destinations, they will "buy more insurance" on rainy days by shifting their departure times earlier in the day. These moment conditions impose that there are no systematic differences in unobserved demand for different *departure times* on rainy versus sunny days. However, since we use the date-demeaned demand shocks here, we allow for the possibility that unobserved demand for the *outside option* is systematically different on rainy days.

D.2.2 Stage 2 moment conditions

In the second stage, we observe HOT $(j = 1)$ but not GP $(j = 0)$ route quantities. Since we can't construct route market shares, we instead match moments of HOT route *quantities*.

HOT intercepts   We match mean HOT quantities in each market. Let q_{1hmt} and $\widehat{q}_{1hmt}(\theta)$ denote the observed and model-predicted HOT quantities, respectively, in departure time h in market m on date t. These moment conditions take the form:

$$\forall m : \qquad \mathbb{E}_{ht}\left[q_{1hmt} - \widehat{q}_{1hmt}(\theta)\right] = 0$$

They are informative about the market-specific HOT intercepts α_{1m}^0.

Mean price and travel time coefficients   We match reduced-form coefficients of HOT quantity on the second-stage price and travel time shifters from Section 6.1.2. First, we match the coefficient on the rounded-up indicator in the price rounding regression discontinuity.[47] Second, we match the coefficients on the GP and HOT crash indicators in column 3 of Table A.3. For each coefficient φ estimated from the data, let $\widehat{\varphi}(\theta)$ denote the analogous coefficient estimated from the model-predicted HOT quantities. These moment conditions impose that $[\varphi - \widehat{\varphi}(\theta)]/\varphi = 0$ for each φ. They are informative about the mean price coefficient $\overline{\alpha}^P$ and the mean travel time coefficient $\overline{\alpha}^D$.

Heterogeneity by income   We match covariances of HOT driver characteristics $\mathbf{x}_{im}$ with attributes of HOT trips taken (price paid and time saved) conditional on the hour and market:

$$\text{Cov}_{ihmt}\left(\mathbf{x}_{im}, \qquad p_{1hmt} \mid q_{i1hmt} > 0, \text{hour}(h), m\right)$$
$$\text{Cov}_{ihmt}\left(\mathbf{x}_{im}, d_{0hmt} - d_{1hmt} \mid q_{i1hmt} > 0, \text{hour}(h), m\right)$$

The data covariances are computed from the toll transaction data, using drivers whom we observe choosing the HOT route at time h on day t. To compute the model covariances, we weight by drivers' predicted probabilities of choosing the HOT route. Matching these moments is informative about the parameters $\mu^{\alpha,P}$ and $\mu^{\alpha,D}$ specifying how preferences for price and travel time vary with driver characteristics.

Unobserved heterogeneity   We match variances and covariances of the same attributes of HOT trips taken (price paid and time saved) conditional on the hour and market:

$$\text{Var}_{ihmt}\left(\qquad p_{1hmt} \mid q_{i1hmt} > 0, \text{hour}(h), m\right)$$
$$\text{Var}_{ihmt}\left(d_{0hmt} - d_{1hmt} \mid q_{i1hmt} > 0, \text{hour}(h), m\right)$$
$$\text{Cov}_{ihmt}\left(p_{1hmt}, \ d_{0hmt} - d_{1hmt} \mid q_{i1hmt} > 0, \text{hour}(h), m\right)$$

---

[47]Specifically, we regress the number of HOT trips on the distance to the price rounding threshold (the running variable), the rounded-up indicator, and (trip definition, hour) fixed effects.

Matching these moments is informative about the variance-covariance matrix $\Sigma^{\alpha,PD}$ of the unobserved component of drivers' preferences for price and travel time.