For Whom the Bridge Tolls: Congestion, Air Pollution, and Second-Best Road Pricing

Matthew Tarduno[†]

September 27, 2021

Job Market Paper

Preliminary and incomplete. Do Not Circulate
Click Here for the Latest Version

Abstract

Cities are increasingly adopting road pricing policies to address the congestion and air pollution externalities associated with urban driving. A first-best road pricing scheme would charge road users for all trips according to the social damages associated with each trip. In practice, road pricing most often takes the form of *cordon zones* — regions in the center of a city where road users are charged for entry. Real-world road pricing schemes therefore deviate from the first-best policy along two important dimensions: First, feasible cordon systems cannot account for all of the heterogeneity in trip-level externalities. Second, cordon zones leave nearby roads unpriced, allowing for externality leakage. As a result, it is generally unclear how to optimally set cordon prices. In this paper, I adapt models from public finance to demonstrate how to optimally set cordon prices in the face of these policy imperfections. Calculating optimal prices requires (i) information about the heterogeneity in marginal trip-level externalities, (ii) the relationship between these externalities and individual price-responsiveness, and (iii) the elasticity of substitution between priced and unpriced trips. I then use detailed microdata from bridge tolls in the San Francisco Bay Area to back out each of these parameters. Armed with this model of urban driving demand, I calculate optimal prices for planned cordon zones in three cities —– San Francisco, Los Angeles, and New York. In each city, I find that leakage drives optimal peak-hour prices ($2-7) well below average social damages ($4-12). As a result, optimal cordon policies are relatively ineffective at internalizing congestion and pollution externalities: In these three cities, second-best cordon prices recover 15 to 40% of the welfare gains that would be achieved under an (infeasible) Pigouvian policy. To conclude, I discuss the prospects for improving the performance of congestion pricing through expanding spatial coverage or allowing for granular time-of-day prices.

[†]University of California at Berkeley; tarduno@berkeley.edu. I thank James Sallee, Michael Anderson, and Reed Walker for support throughout this project. I also thank Lucas Davis, Meredith Fowlie, Marco Gonzalez-Navarro, Alan Auerbach, Matthew Gibson, Josh Apte, Max Auffhammer, Sofia Villas-Boas, James Sears, Jenya Kahn-Lang, Marshall Blundell, Matthew Suandi, and seminar participants at NC State's Camp Resources for their helpful feedback. Lastly, I would like to thank Jeff Gerbracht for invaluable assistance accessing and understanding the tolling data used in this paper.

Contents

1. Introduction

Economists have long advocated for charging road users to address the negative externalities associated with urban driving (Vickrey, 1963; Johnson, 1964; Parry, 2002). Following early policy experiments in Singapore and London, a growing number of cities including New York, Los Angeles, and San Francisco, are considering implementing road pricing. Despite its history of advocating for road pricing, however, the economics literature offers relatively little on how to implement road pricing in practice: If city policymakers decide to price roads, how should they set these prices? And how should they adjust pricing schemes to address the imperfections that come with real-world policy?

A first-best road pricing policy would charge drivers for the marginal social damages (the time cost imposed on others plus the social cost of pollution generated) associated with every vehicle trip. Practical constraints, however, render first-best road pricing infeasible in most settings. Implementing a first-best policy would require detailed information about each driver's routes and emissions, as well as real-time traffic data. It is typically too costly to collect this information through a passive sensor network, and proposals for GPS-based pricing schemes are typically rejected on privacy grounds (Lehe, 2019; Giuliano, 1992). Consequently, city-wide road pricing most often takes the form of *cordon zones* — regions in the center of a city where drivers are charged for entry. Real-world road pricing schemes therefore deviate from the first-best policy along two important dimensions: First, feasible cordon systems cannot account for all of the heterogeneity in trip-level externalities. Second, cordon zones leave nearby roads unpriced, allowing for externality leakage. As a result, it is generally unclear how to set cordon prices even if policymakers have perfect information about the social damages associated with trips that pass through the city center (Parry, 2009).

In this paper, I adapt models from public finance to characterize optimal cordon prices in the face of these policy imperfections. Calculating optimal prices requires (i) information about the heterogeneity in marginal trip-level externalities, (ii) the correlation between trip-level externalities and individual price-responsiveness, and (iii) the elasticity of substitution between priced and unpriced trips. Outside of road pricing, this framework can be applied to any setting where externality heterogeneity and leakage simultaneously prevent the implementation of a first-best corrective policy (e.g., electricity markets, or sin taxes).

In the empirical section of this paper, I leverage a natural experiment from the San Francisco Bay Area to recover estimates of each of the parameters necessary to calculate optimal cordon prices. In 2010, bridge tolls increased on all of the region's bridges, and peak-hour pricing was implemented on the region's busiest bridge. I use this variation in road prices together with detailed electronic tolling microdata to estimate a model of traffic demand. The results from this exercise imply that the two policy imperfections — leakage and heterogeneity — create a tension in optimal cordon pricing. Trips associated with higher social damages are more elastic. Absent leakage, this heterogeneity would imply second-best optimal prices that are *above* average social damages (Diamond, 1973). The discrete spatial and temporal cutoffs in cordon pricing, however, incentivize some drivers to shift trips in time and space to avoid tolls. Absent heterogeneity, this leakage would imply optimal prices that are *below* average social damages (Green and Sheshinski, 1976).

I use this model of travel demand to calculate second-best optimal prices for the proposed cordon zones in San Francisco, Los Angeles, and New York. I find that the leakage effect strongly dominates the heterogeneity effect in each of these cities. In San Francisco, for example, when cordon prices are constrained to peak hours the second-best optimal prices that account for both heterogeneity and leakage are \$2 to \$3. This is roughly half of the average social damages generated by trips that use the cordon during those periods (\$4 to \$6). Unsurprisingly, this policy performs poorly relative to the (infeasible) Pigouvian prescription. The second-best optimal road pricing scheme in San Francisco achieves only 28% of the total welfare gains, 30% of the congestion reductions, and 22% of pollution reductions relative to a policy where drivers are charged according to the marginal damages of each trip. Across the three cities that I examine, I find that optimal peak-hour

cordon prices are more effective at internalizing congestion than they are at internalizing pollution. This reflects the fact that while congestion and pollution externalities are spatially correlated, average trip-level pollution damages do not exhibit the same within-day variation as congestion externalities.

To conclude, I investigate the prospects for improving cordon pricing policies. Allowing for flexible hourly prices between 6 a.m. and 7 p.m. generates sizable welfare gains: $146 million in San Francisco, $157 million in Los Angeles, and $286 million in New York. In each city, however, a flexible hourly cordon price would still leave a majority of the possible welfare gains unrealized.

This paper makes three primary contributions:

First, this paper provides the first empirical estimates of optimal cordon prices that account for both pollution and congestion. I recover optimal peak-hour cordon prices that range from $2.10 in San Francisco to $7.92 in New York. While there are robust literatures documenting the reduced-form relationship between road pricing and traffic speeds (Yang, Purevjav, and Li, 2020; Gibson and Carnovale, 2015; Leape, 2006) as well as traffic and local air pollution (Currie and Walker, 2011; Anderson, 2020; Gibson and Carnovale, 2015; Knittel, Miller, and Sanders, 2016; Tonne, Beevers, Armstrong, Kelly, and Wilkinson, 2008), these results have yet to be combined into optimal cordon prices that account for both of these externalities (Parry, 2009). Importantly, the optimal road prices presented in the paper also account for imperfections in real-world policies. Both theoretical and empirical studies suggest that while price or quantity-based cordons can ameliorate pollution and congestion in some settings (Zhong, Cao, and Wang, 2017; Börjesson, Eliasson, Hugosson, and Brundell-Freij, 2012), policies designed without regard to agent re-optimization and heterogeneity may lead to poor or perverse policy outcomes (Davis, 2008, 2017; Zhang, Lawell, and Umanskaya, 2017; Hanna, Kreindler, and Olken, 2017; Green, Heywood, and Paniagua, 2020). As such, this paper explicitly calculates optimal cordon prices through a second-best tax framework.

Second, this paper contributes to the literature on externality taxation by characterizing second-best prices in the presence of both heterogeneous externalities *and* externality leakage. This framework extends two canonical models of second-best pricing: the "Diamond" model (Diamond (1973), see also Knittel and Sandler (2018)), which shows that second-best uniform prices are a weighted average of heterogeneous externalities, and the "leakage" model, where second-best optimal prices reflect marginal damages, less a term that captures leakage (substitution) to other unpriced goods that also generate externalities (Green and Sheshinski (1976), see also Davis and Sallee (2020); Gibson (2019); Holland (2012)). Specifically, I consider the setting where there are many externality-generating goods, externalities vary across consumers and goods, and only a subset of the goods are taxable. I show that in the presence of both heterogeneity and substitution, the optimal second-best tax formula combines characteristics of the canonical Diamond and leakage models. Holding fixed all other taxes, the optimal tax on any *one* good is the Diamond-weighted marginal damages associated with the consumption of the good, less a term governed by the Diamond-weighted leakage to other goods. The optimal second-best tax vector solves a system of equations, where terms in this system reflect individual externalities, own-price elasticities, and cross-price elasticities. This characterization is most closely related to Allcott, Lockwood, and Taubinsky (2019), who characterize the optimal vector of taxes on sugary drinks in the setting with welfare weights that reflect a planner's distaste for inequality. The optimal taxation problem in this paper also resembles the optimal collection of government revenue when taxes are distortionary (Ramsey, 1927). Namely, the solution to the road pricing problem involves a matrix of substitution elasticities, as does the general solution to the canonical Ramsey tax problem. In this setting, an untaxed good's idiosyncratic externality is analogous to each good's distortions in the many-good Ramsey problem.

This extension of optimal second-best pricing is applicable in settings outside of transportation. In electricity markets, for example, the externalities associated with generation differ based on the location of plants (urban

or rural, upwind or downwind of population centers), and policies implemented by states or utilities may allow for externality leakage if electricity is imported from other jurisdictions. Sin taxes (e.g., alcohol taxes, cigarette taxes) similarly have heterogeneous impacts on consumers, and taxing any single product may induce consumers to substitute towards related (and undertaxed) sin products.

Lastly, this paper presents a new approach for estimating the willingness of commuters to shift the schedule of their trips. Scheduling costs are a key determinant of urban congestion Kreindler (2018), and a primitive parameter in the canonical "bottleneck" model of urban travel (Vickrey (1963), Arnott, De Palma, and Lindsey (1990), Arnott, De Palma, and Lindsey (1993)). Adapting tools from the public finance literature on bunching (Saez, 2010; Kleven and Waseem, 2013), I develop an estimator that infers scheduling costs from the excess density of trips taken during times of day that fall just outside a peak pricing window. Because peak pricing is used to alleviate congestion in bridges and tunnels in many cities, this estimation approach can be applied in most metro areas.

The rest of this paper is organized as follows. Section 2 characterizes the second-best optimal externality taxes in the presence of heterogeneity and leakage. Section 3 details the structural model of travel demand that I use to back out the statistics necessary to estimate optimal prices. Section 4 outlines the setting and natural experiment that I use to estimate the model of travel demand, and Section 5 covers the data. In Section 6, I describe the empirical strategy that I use to estimate the model of travel demand. I present results in Sections 7 and 8, discuss these results in Section 9, and conclude in Section 10.

2. Theory: Externality Taxation Under Heterogeneity and Substitution

Public economics provides an unambiguous prescription for addressing market externalities: apply a (Pigouvian) tax equal to the marginal damages associated with consuming the externality-generating good. In practice, however, policy instruments typically lack the precision and coverage to execute this prescription. When corrective taxation cannot account for heterogeneous externalities or substitution to other externality-generating goods, the (second-best) optimal tax on any given good may differ substantially from the tax instituted in the ideal Pigouvian policy. In this section I outline canonical models for optimal taxation under each of these separate imperfections (heterogeneity and substitution), and then present a model that can be applied to instances where heterogeneity and substitution simultaneously prevent the implementation of the first-best.

2.1. Heterogeneity

Diamond (1973) characterizes the optimal taxation of a good which generates heterogeneous externalities when consumed by different agents. The single (uniform) price cannot perfectly correct for all externalities. Diamond shows that when demand elasticities are also heterogeneous, the optimal second-best tax on the externality-generating good is a weighted average of the individual externalities, where the weights (henceforth *Diamond weights*) are the individual own-price elasticities.

Formally, consider n consumers that derive utility from their consumption of an externality-generating good, α_i, and disutility from other's consumption of this good:

$$U^i = U(\alpha_1, ..., \alpha_i, ...\alpha_n) + \mu_i$$

The second-best optimal uniform tax in this setting is:

$$\tau^* = \frac{-\sum_h \sum_{h \neq i} \frac{\partial U^h}{\partial \alpha_i} \alpha_i'}{\sum_h \alpha_h'} \tag{1}$$

Where α_i' is the derivative of consumer i's demand for α with respect to the price of α, and $\frac{\partial U^h}{\partial \alpha_i}$ is the marginal external cost that consumer i imposes on consumer h by consuming α.

This expression captures an important principle in second-best corrective taxation: If individual elasticities are positively (negatively) correlated with idiosyncratic externalities, the second-best uniform tax on the externality-generating good will be larger (smaller) than the naive average of marginal damages. Intuitively, the role of corrective taxes is to move individuals to adjust their consumption of a product to the level where private marginal benefit equals the social marginal cost. If a given group is unresponsive to price, however, the second-best optimal tax described above will provide the correct incentive for the responsive group to consume at the level that balances private and social marginal costs.

In the setting of cordon prices, heterogeneity in externalities may arise from several factors. Sources of congestion heterogeneity include the length of the trip, the time that the trip is taken, and the specific roads used within and outside of the cordon. Sources of pollution heterogeneity include vehicle attributes, travel speed[1], and trip length.

2.2. Substitution

In many cases, legal, political, or practical constraints prevent policymakers from pricing all related externality-producing goods. In the case of two externality-generating goods (one of which is taxable and one of which is not) and homogeneous marginal damages, the second-best prescription is to tax at the marginal damages of the first good, less a term that is increasing in the substitutability of the two goods, and increasing in the marginal damages of consuming the untaxable good.

Formally, consider two goods, x and y, with associated marginal external damages ϕ_x and ϕ_y. A representative consumer with an exogenous income derives utility from these two goods, and a quasilinear numeraire good:

$$U = U(x, y) + z$$

If a social planner is constrained to only tax x, the optimal tax is:

$$\tau_x^* = \phi_x + \frac{dy/dp_x}{dx/dp_x}\phi_y \tag{2}$$

This formula is used in Holland (2012), Davis and Sallee (2020), and reflects separability insights from Kopczuk (2003). In the remainder of this section, I cover two extensions to the above models. In section 2.3, I characterize optimal taxes for a general set of externality-generating goods, where only a subset of them can be taxed. In section 2.4, I characterize optimal taxes for a general set of externality-generating goods, where only a subset of them can be taxed, *and* marginal damages are heterogeneous by consumer.

2.3. Substitution with Many Goods

In this section, I extend the two-good model in section 2.2 to the case of many externality-generating goods, some of which are untaxable. This problem is a generalization of the two and three-good direct vs. indirect

[1] Verhoef, Nijkamp, and Rietveld (1995) provide a theoretical overview of the Diamond model as it applies to congestion and pollution externalities.

taxation problems presented in Green and Sheshinski (1976) and Sandmo (1978).

Setup: A representative consumer chooses quantities of M goods, $(h_1, ..., h_M)$ and a numeraire, z. Each non-numeraire good has an associated externality, ϕ_m. A policymaker can choose tax levels τ_j for goods $j \in [1, J]$ where $J < M$. I assume goods $k \notin [1, J]$ are un- or under-taxed.

In Appendix A, I show that under these constraints the optimal tax for good j holding fixed the taxes on all other taxable goods k is:

$$\tau_j = \phi_j + \frac{1}{\frac{\partial h_j}{\partial p_j}}(\sum_{k \neq j}^{J} \frac{\partial h_k}{\partial p_j}[\phi_k - \tau_k] + \sum_{l=J+1}^{M} \frac{\partial h_l}{\partial p_j}\phi_l) \tag{3}$$

This intermediate results is intuitive. Holding fixed all taxes other than τ_j, the optimal value for this final tax is its externality, ϕ_m, plus a term that captures the extent to which consumers switch to other goods, and the level of unpriced externality of those goods. Identifying the optimal tax vector requires simultaneously solving J equations in the form of Equation 3.

To do so, one can rewrite Equation 3 to separate the tax and externality terms:

$$\tau_j + \frac{1}{\frac{\partial h_j}{\partial p_j}}(\sum_{k \neq j}^{J} \frac{\partial h_k}{\partial p_j}\tau_k) = \phi_j + \frac{1}{\frac{\partial h_j}{\partial p_j}} \sum_{l=1}^{M} \frac{\partial h_l}{\partial p_j}\phi_l$$

This yields J equations, each linear in the J tax levels:

$$\alpha_1^j \tau_1 + ... + \alpha_k^j \tau_k + ... + \alpha_J^j \tau_J = b_j \qquad \forall j \in [1, J] \tag{4}$$

Where α_k^j and b_m are defined as:

$$\alpha_k^j = \frac{\frac{\partial h_k}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \tag{5}$$

$$b_j = \phi_j + \sum_{l=1}^{M} \frac{\frac{\partial h_l}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}}\phi_l \tag{6}$$

The α and b terms have intuitive interpretations. α_k^j is the share of the reduction in overall consumption of good j that shifts to good m as a results of an increase in the price of good j. b_j is the overall reduction in externalities that results form the increase in the price of good j; this consists of a direct component, ϕ_j, plus a (negative) leakage component, $\sum_{l=1}^{M} \frac{\partial h_l}{\partial p_j} / \frac{\partial h_j}{\partial p_j}\phi_l$.

This system can be written compactly as:

$$\begin{bmatrix} \alpha_1^1 & ... & \alpha_J^1 \\ & ... & \\ \alpha_1^J & ... & \alpha_J^J \end{bmatrix} \begin{bmatrix} \tau_1^* \\ ... \\ \tau_J^* \end{bmatrix} = \begin{bmatrix} b_1 \\ ... \\ b_J \end{bmatrix}$$

$$\boldsymbol{A\tau} = \boldsymbol{b} \tag{7}$$

The optimal tax vector when there are J taxable goods out of M total externality-generating goods is:

$$\boldsymbol{\tau} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{8}$$

Equation shows that solving for the second-best optimal vector of corrective taxes in a setting with incomplete tax coverage and substitution between many externality-generating goods requires a) the consumption externalities associated with each good, and b) the substitution matrix between all goods.[2]

2.4. Heterogeneity and Substitution

Finally, I consider the case where a) there are many externality-generating products, b) policymakers can tax only a subset of these products, and c) externalities are heterogeneous in consumption of the products.

While I apply this model to urban driving externalities in this paper, many markets feature externalities and policy instruments that fit this description. Electricity generation, for example, produces environmental externalities that vary by location (Muller and Mendelsohn, 2007; Hernandez-Cortes and Meng, 2020), and local environmental policies may induce leakage if utilities import electricity across jurisdictional borders. Similarly, the consumption of "sin" goods may be associated with externalities or internalities that vary across consumers, and taxing any one product (e.g., cigarettes) may induce leakage towards other products (e.g., vape pens) that do not fall under a policymaker's purview.

Setup: N Heterogeneous consumers choose between M externality-generating goods and a numeraire, z. I denote individual i's consumption of good m as h_i^m. Each individual has an exogenous income μ_i. I assume that each consumer's utility is a function of their consumption of these M goods and a quasilinear numeraire, as well as other's consumption of these goods (which generate externalities and decrease i's utility): $U_i(h_1^1, ...h_1^M, ..., h_i^1, ...h_i^M, ...h_N^1, ...h_N^M) + z_i$.

As in section 2.3, a policymaker can choose tax levels for goods $j \in [1, J]$ where $J < M$. I assume goods $m \notin [1, J]$ are un- or under-taxed. I denote τ^j as the tax on good j.

In Appendix A, I show that the optimal tax on τ_j as a function of the k other tax levels is:

$$\tau_j = \frac{\sum_{i=1}^{N}\sum_{g}^{N}\left(\frac{\partial U^i}{\partial h_g^1}\frac{\partial h_g^1}{\partial p_j} + ... + \frac{\partial U^i}{\partial h_g^M}\frac{\partial h_g^M}{\partial p_j}\right)}{\sum_{i=1}^{N}\frac{\partial h_i^j}{\partial p_j}} + \frac{\sum_{k\neq j}^{J}\frac{\partial h_i^k}{\partial p_j}\tau_k}{\sum_{i=1}^{N}\frac{\partial h_i^j}{\partial p_j}} \tag{9}$$

This expression for the optimal level of a given tax is equivalent to the equation for substitutes with homogeneous damages where each of the marginal damages have been replaced by Diamond-weighted externalities that account for heterogeneity in marginal damages across individuals. As in the case of many substitutes with homogeneous damages, the optimal tax vector solves a system of J equations:

$$\begin{bmatrix} \alpha_1^1 & ... & \alpha_J^1 \\ & ... & \\ \alpha_1^J & ... & \alpha_J^J \end{bmatrix} \begin{bmatrix} \tau_1^* \\ ... \\ \tau_J^* \end{bmatrix} = \begin{bmatrix} b_1 \\ ... \\ b_J \end{bmatrix}$$

$$\boldsymbol{A\tau} = \boldsymbol{b} \tag{10}$$

[2]Note that this substitution matrix contains cross-price consumption *derivatives* and not cross-price consumption *elasticities*. $\boldsymbol{A}$ contains 1's along the diagonal; off-diagonal terms fall in the closed interval $[0, -1]$.

Where α_k^j and b_k are defined as:

$$\alpha_k^j = \frac{\sum_{i=1}^{N} \frac{\partial h_i^k}{p^j}}{\sum_{i=1}^{N} \frac{\partial h_i^j}{\partial p_j}} \tag{11}$$

$$b_j = \underbrace{\frac{\sum_i^N \sum_{g \neq i}^N \frac{\partial U_i}{\partial h_g^j} \frac{\partial h_g^j}{\partial p_j}}{\sum_i^N \frac{\partial h_i^j}{\partial p_j}}}_{\text{Diamond-weighted externality of good } j} + \underbrace{\sum_{l \neq j}^{M} \frac{\sum_i^N \sum_{g \neq i}^N \frac{\partial U_i}{\partial h_g^l} \frac{\partial h_g^l}{\partial p_j}}{\sum_l^N \frac{\partial h_i^j}{\partial p_j}}}_{\text{Diamond-weighted leakage shares}} \tag{12}$$

Solving for the second-best optimal vector of corrective taxes therefore requires (i) the (heterogeneous) externalities associated with each good, (ii) the relationship between these heterogeneous externalities and individual price elasticities, and (iii) individual-level substitution matrices between goods. In what follows, I recover these parameters in these setting of urban driving demand, and apply this framework to recover optimal cordon prices in three US Cities.

3. A Structural Model of Driver Behavior

The theory outlined in Section 2 implies that calculating the second-best optimal cordon prices requires information about the heterogeneity in the price responsiveness of different types of trips that cross a cordon, as well as the rates of substitution between trips that can and trips that cannot be priced. To recover these parameters, I estimate a canonical "bottleneck" model of driving demand (Arnott, De Palma, and Lindsey, 1990, 1993).

Formally, imagine drivers i who choose between departure times h and a routes r to satisfy their demand for travel. Included in this choice set is the outside (no trip) option, which is normalized to zero utility. Each driver has an exogenous ideal arrival time, h_i^A. Drivers are atomistic and face travel times $T(h,r)$ and tolls $p(h,r)$ that may vary by route and time of day. A driver arriving before or after her ideal arrival time incurs disutilities γ_e and γ_l per minute, respectively. Drivers also receive disutility α from each minute spend commuting. Utility is thus:

$$u(h_i, r_i) = -\alpha T(h_i, r_i) - \gamma_e \underbrace{|h_i + T(h_i, r_i) - h_i^A|_-}_{\text{time early}} - \gamma_l \underbrace{|h_i + T(h, r) - h_i^A|_+}_{\text{time late}} - \beta p(h_i, r_i) \tag{13}$$

Each driver chooses the route (r_i) and time of day (h_i) that maximizes their expected utility:

$$\{h_i^*, r_i^*\} \in arg\ max_{h_i,}$$

To clarify the mapping between this structural model and the optimal tax formula (Equation 10), a "good" is a trip taken on a given *route* at a given *time of day*. Typical cordon zones have discrete spatial and temporal cutoffs.[3] "Substitution" refers to drivers adjusting these trips in time (h) and space (r) to avoid tolls. "Heterogeneity" refers to the fact that trips that cross a city cordon zone during a given time period may differ in pollution externalities (a function of trip length, vehicle characteristics, and travel speed) as well as congestion externalities (a function of trip length and traffic density along the trip).

[3]The London Cordon Zone, for example, charges road users £15 between 7 a.m. and 10 p.m. for entering the city center. The Milan Cordon Zone charges users €2 to €5 based on vehicle type between 7:30 am and 7:30 pm. The proposed cordon zones in both New York and San Francisco exhibit similar sharp spatial and temporal cutoffs.

The estimated parameters of Equation 13 imply a matrix of own and cross-price elasticities between routes and hours of day that I use to solve for second-best cordon prices using the framework outlined in Section 2.

4. Natural Experiment: Traffic Tolling in the San Francisco Bay Area

I use detailed tolling data from the San Francisco Bay Area together with revisions to bridge toll prices between 2010 and 2012 to estimate the model of travel outlined in Section 3.

4.1. Bay Area Bridge Tolls

FasTrak is the electronic toll system used in California. Drivers are charged for using certain roads (bridges and high-occupancy toll lanes) via transponders mounted to the car's dash.[4] In the San Francisco Bay Area, tolls are collected on each of the region's trans-bay bridges (mapped in Figure 1) for westbound trips only.

Figure 1 — San Francisco Bay Area Bridges

Figure 1: This maps shows the four San Francisco Bay Area bridges covered in this study. The *Richmond Bridge* connects Richmond California and the eastern Bay Area to San Rafael California and Marin County. The *Bay Bridge* connects Oakland California to San Francisco, California. The *San Mateo Bridge* connects Hayward California to San Mateo, California. The *Dumbarton Bridge* connects Fremont, California to Palo Alto, California. Each of these bridges charges drivers for westbound trips.

[4]Drivers can pay with cash if they do not purchase a FasTrak device. Between 2010 and 2019, cash payers represented roughly 10% of all trips on Bay Area bridges.

4.2. Variation in Toll Prices 2010 - 2012

Bay Area FasTrak tolls vary by bridge, vehicle type, and time of day. I focus on passenger vehicles (as opposed to light and heavy-duty trucks), which constitute roughly 97% of vehicle trips on Bay Area bridges.[5] Currently, passenger vehicles are charged between $3 and $7 dollars, depending on the time of day, the number of occupants, and whether or not the vehicle is electric/hybrid.

In this paper, I leverage several changes in the tolling structure that occurred on July 1st, 2010 to identify the parameters necessary to calculate optimal road prices. In 2009, the Bay Area Toll Authority (BATA) adopted a Resolution 90, which increased the base prices for passenger vehicles from $4 to $5 beginning on July 1st, 2010, and established peak-hour pricing (detailed below). The current toll rates reflect a second change to bridge prices that followed Regional Measure 3, a 2018 ballot initiative that increased tolls for all vehicle types by $1 beginning on January 1st, 2019. This intertemporal variation in toll prices is plotted in Figure 2.

Figure 2 — Variation in Passenger Vehicle Bridge Tolls

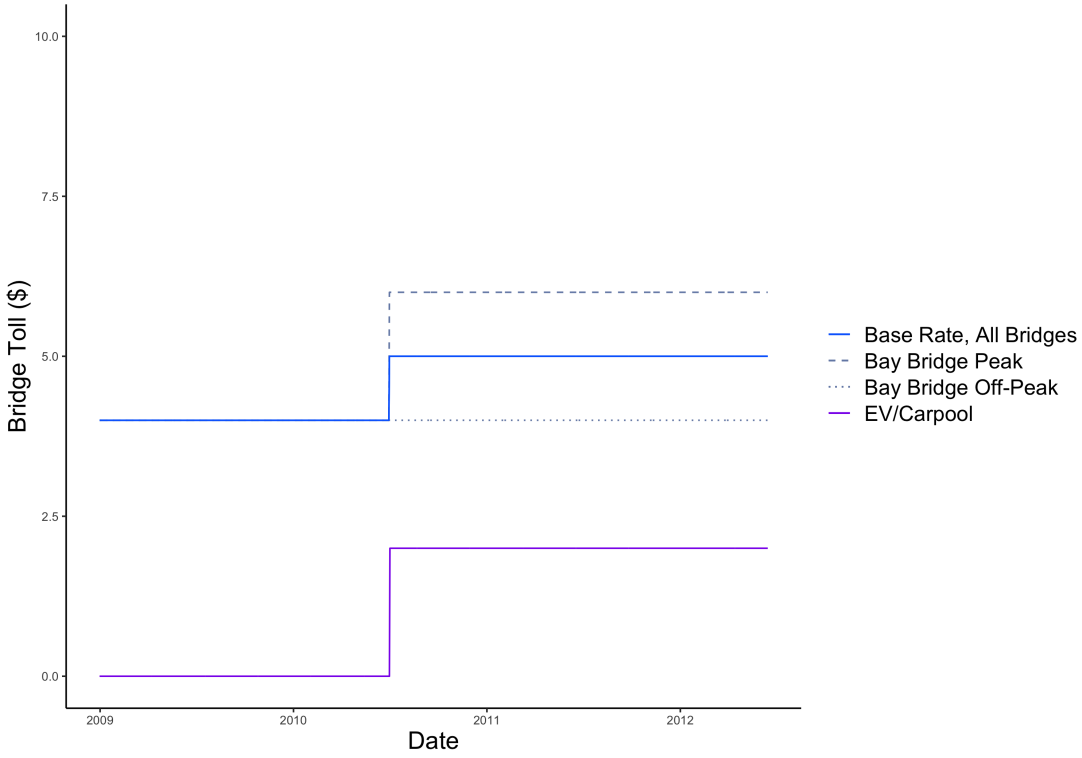


Figure 2: This figure shows Bay Area bridge tolls between 2009 and 2012 for passenger vehicles. Prices are uniform across bridges, with the exception of the *Bay Bridge*, which connects San Francisco and Oakland. Beginning in 2010, passenger vehicles crossing the Bay Bridge faced a two-dollar difference between peak and off-peak prices. The peak and off-peak prices are plotted above as dotted and dashed lines, respectively. EV and carpool trips were free prior to 2010. Beginning in July of 2010, EV/carpool trips were charged the base rate ($5), except during peak hours, where they receive a discount ($2.5) on all bridges.

4.3. Peak-hour Pricing on the Bay Bridge

To address acute congestion on the region's busiest bridge, the Bay Area Toll Authority imposed peak hour pricing on the Bay Bridge beginning on July 1st 2010. Passenger vehicles crossing westbound through the Bay

[5]Between 2009 and 2019, the four major Bay Area bridges recorded roughly 285,000 FasTrak transactions daily for passenger vehicles, versus 7,000 daily transactions for vehicles with three of more axles.

Bridge toll plaza on weekdays between 5 a.m. and 10 a.m., or between 3 p.m. and 7 p.m. (henceforth *peak hours*) were charged $6. Tolls for all other hours (*off-peak*) remained at the pre-2010 price of $4.

Prior to July 1st, 2010, passenger vehicles with two or more passengers, as well as eligible electric and hybrid electric vehicles were not subject to tolls on any Bay Area bridges. Starting in 2010, these vehicles were subject to the full toll value during off-peak hours, but retained a discount during peak hours: EV/carpool trips were charged $2.50 to use Bay Area bridges between July 1st, 2010 and January 1st, 2019.

Figure 3 — Peak Pricing on Bay Area Bridges

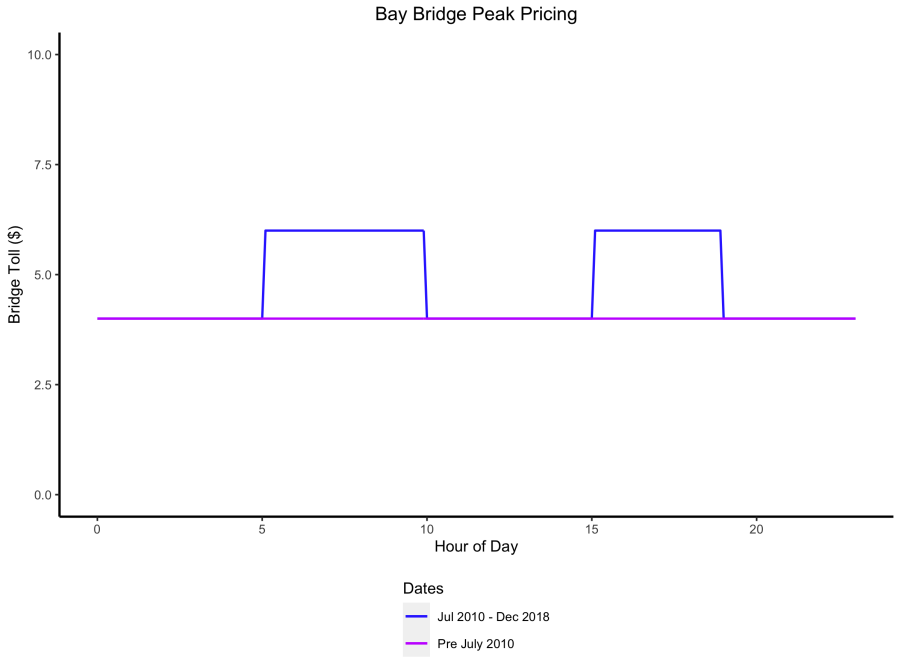


Figure 3: This figure displays peak-hour pricing schemes for passenger vehicles (vehicles with two axles) on California's *Bay Bridge*, which connects San Francisco and Oakland. Beginning on July 1st, 2010, passenger vehicles crossing westbound on weekdays during peak hours (5 a.m. and 10 a.m., or between 3 p.m. and 7 p.m.) faced higher prices than vehicles crossing during off-peak hours. Peak-hour prices are displayed on large variable-message sign about the Bay Bridge toll plaza. Weekend trips on the Bay Bridge and trips on the other major Bay Area bridges are not subject to peak pricing, instead charcing the the base rate for passenger vehicles ($4 for pre-2010 and $5 for July 2010 - December 2018).

5. Data

5.1. Reconstructing Choice Sets

Estimating the model outlined in Section 3 requires individual-level data on travel choices, travel times, and road prices. To construct this choice set, I combine detailed microdata from the FasTrak tolling system with historic travel time data purchased from TomTom's Historic Traffic Stats database.

FasTrak Toll Data: I use microdata from the FasTrak tolling system to create a panel of individual-level travel decisions. These microdata record any transactions that occurred on the four trans-bay bridges between January 1st 2009 and July 1st 2019. A single observation in this data set includes the date, time, and location of the vehicle crossing, as well as the vehicle class, the price paid, and an indicator for whether the vehicle used the EV/carpool lane. For vehicles with registered FasTrak devices (vehicles that did not pay cash) the microdata also include a unique FasTrak id number, as well as a zip code associated with the FasTrak holder. These data contain hundreds of millions of trip records. I restrict the dataset on several dimensions. First, I restrict the dataset to devices with a valid (Bay Area) zip code, as this information is necessary for estimating travel times. Second, I drop devices with infrequent use (fewer than 50 weekday trips in the year prior to the 2010 price change), or users that take multiple trans-bay trips per day (greater than 500 weekday trips in the year to the 2010 price change). Lastly, for the purposes of estimation, I consider only trips taken in a narrow window (weekdays between June 15th to July 15th) before and after the 2010 change in toll prices. The resulting panel consists of 32,104 individuals and 1,078,044 bridge crossings.

Travel Time Data: Because the FasTrak microdata include only the device zip code and bridge used, I must infer trip travel times. I do so in two steps.

First, based on the zip code and travel behavior of a given vehicle, I use data from the 2012 California Household Travel Survey (CHTS) to infer a probability distribution over destinations for that vehicle. For example, if I observe a driver from Oakland traveling via the Bay Bridge, I enumerate the destination cities of all CHTS drivers from Oakland who reported using the Bay Bridge. I repeat this for all of the driver's trips, resulting in a probability distribution over endpoints for each FasTrak device.

Second, I use TomTom's historic data to reconstruct the travel time between an individual's home zip code and each of the possible destination endpoints. The FasTrak data provide hourly traffic speeds for major roads in the year before and the year after the July 2010 adjustment to Bay Area tolls. Importantly, I also use the TomTom data to estimate counterfactual travel times. The result is a recreation of each driver's choice set, namely the travel time and price for each trip that driver took, as well as the price and travel time if they had taken that same trip at a different time of day, or using a different bridge. This choice set construction is described in full detail in Appendix D.

Ideal Arrival Times: Ideal arrival times, h_i^A, are not directly observed, and therefore must be inferred form each driver's activity. For each driver, I assign h_i^A as the modal bridge crossing time of each individual during weekdays between January 1st, 2010 and July 1st, 2010.

5.2. Externalities

Although data on trip-level externalities is not necessary for estimating a model of travel demand, second-best optimal road prices depend on the correlation between the price elasticity of demand for a given trip and the idiosyncratic externalities associated with that trip (see Section 2). I therefore estimate the externalities (congestion and pollution) associated with each FasTrak trip. Note that I do not include accident externalities when calculating trip-level externalities. I provide a detailed rationale for excluding accident externalities in Appendix F.

Congestion Externalities: Congestion externalities are highly space and time dependent. The transportation economics literature canonically presents congestion externalities as a function of traffic *density*, measured in vehicles per lane-mile (Small, Verhoef, and Lindsey, 2007). To assign congestion externalities to trips in the FasTrak dataset, I use estimates from Yang, Purevjav, and Li (2020), who show that the marginal external (travel time) cost of traffic is convex in traffic density. That is, congestion externalities are negligible when there are few other vehicles on the road, but increase sharply with the number of vehicles per lane-mile. The congestion costs from this paper are reproduced in Figure 4.

Using a comprehensive network of traffic sensors on roadways in the Bay Area I infer the density along the route for each FasTrak trip. For each trip, I use HERE Technology's *Routes* API to identify the likely route between the zip code associated with the device and the bridge crossed. For each traffic sensor along the driver's route, I use estimates from Yang, Purevjav, and Li (2020) to assign a marginal external congestion cost (in dollars per mile)[6] to this point based on the average traffic density at that sensor at the time of day when the trip was taken. A trip's total congestion externality is then the average of the external congestion costs (in dollars per mile) along the route times the length of the trip.

As noted above, because one of the trip termini is missing from the FasTrak data, I impute the congestion externalities for the missing segment of the trip (between the bridge to the place of work) using the likely destination locations conditional on observable characteristics (home zip code, bridge used, time of day). Note that the majority of variation in externalities is driven by the choice of bridge and time of day, suggesting any noise in this imputation process should not meaningfully impact the results in this paper.

FIGURE 4 — CONGESTION COSTS, REPRODUCED FROM YANG ET AL. (2020)

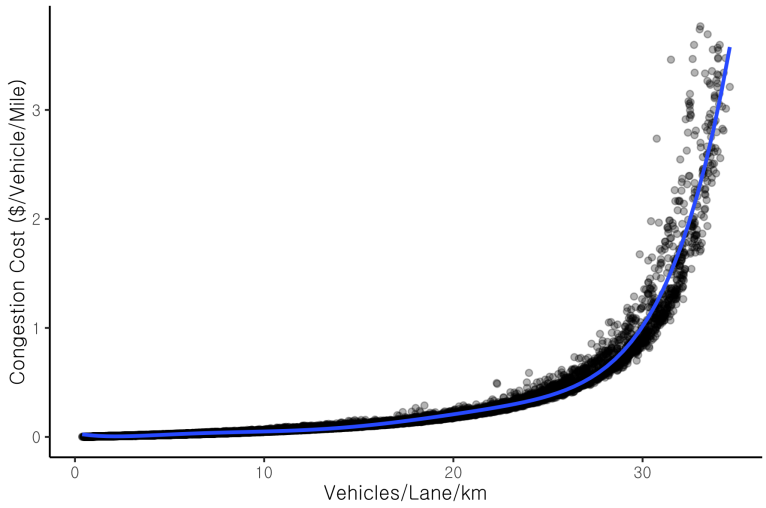


Figure 4: Congestion costs estimated by Yang, Purevjav, and Li (2020), who exploit variation in traffic density imposed by Beijing's driving restriction policy. The original results are presented in Yuan/Vehicle/km. I convert these values to dollars by a) converting currencies, and b) replacing a the Beijing-specific value of time from (50% of the average wage rate in Beijing) with a $20 value of travel time.

Emissions Externalities: Fuel combustion and brake wear in passenger vehicles generates several air pollutants. These include "global" pollutants like CO_2 and methane, which contribute to climate change, as well as "local" pollutants like particulate matter (PM), nitrogen oxides (NO_x) and reactive organic compounds (ROCs),

[6]The estimates from Yang, Purevjav, and Li (2020) are in yuan/vehicle/km. I convert these values to dollars by a) converting currencies, and b) replacing a the Beijing-specific value of time from (50% of the average wage rate in Beijing) with a $20 value of travel time.

which negatively impact the health of nearby residents (Anderson, 2020; Currie and Walker, 2011; Deryugina, Heutel, Miller, Molitor, and Reif, 2019). Vehicle emissions factors — the amount of a particular pollutant that a vehicle emits while traveling a mile — depends on a number of variables, including the type of fuel consumed, the fuel economy, the vehicle vintage[7], and vehicle speed.[8]

I estimate emissions for FasTrak trips using data from the California Air Resource Board's Emissions Factor Database (EMFAC). This database contains estimates of the average emissions rates of vehicles registered in each county as a function of vehicle speed. I then assign social costs to these trip-level emissions. I value global pollutants using the EPA's 2021 social cost of carbon ($51 per ton) and methane ($1,500), respectively. Local pollutant damages reflect the cost of emitting each pollutant at ground level in San Francisco, according the EASIUR model of local pollution damages. See Appendix B for details on individual pollutant costs.

Figure Figure 5 — Pollution Externalities at Various Speeds

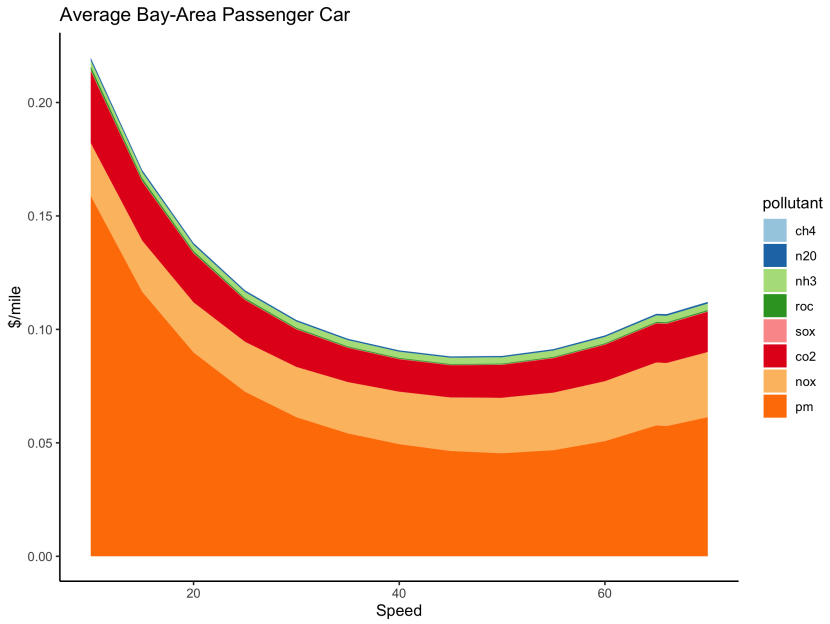


Figure 5: This figure shows per-mile pollution externalities for the average passenger vehicle in the Bay Area. These costs reflect fleet-average emissions factors (in grams/mile) of different pollutants at different speeds reported by California's Emissions Factor Model (EMFAC), multiplied by the social cost of each pollutant. For local pollutants, the social cost is calculated using the Estimating Air pollution Social Impact Using Regression (EASIUR) Online Tool, calibrated with coordinates from San Francisco. For global pollutants, I use the EPA's 2021 social costs of $51 per ton of CO_2 and $1,500 per ton of CH_4, respectively. All values are in 2020 dollars.

Together, the data described in this section allow me to recreate the choices and choice set facing a sample of Bay Area drivers, augmented with estimates of the social costs associated with each trip choice. In what follows, I use these data to recover the parameters necessary for calculating optimal prices for San Francisco's cordon zone.

[7]Older vehicles have higher emissions factors for two reasons: They were subject to less stringent tailpipe emissions and fuel economy standards when they were built, and emissions abatement technologies (catalytic converters) depreciate over a vehicles lifetime

[8]Vehicle speed impacts emissions through engine efficiency and the intensity of brake ware.

6. Empirical Strategy

I use two strategies to recover the primitives that determine driving behavior. In my preferred specification, I use the variation in toll prices in 2010 together with the FasTrak microdata to estimate the parameters of Equation 13 using a miltinomial logit model. As a check for the results from this first method, I apply a bunching estimator to the Bay Bridge's notched tolling schedule, producing a second set of empirical estimates of scheduling costs.

6.1. Multinomial and Mixed Logit

As described in Section 5, the FasTrak microdata and the TomTom historic traffic data allow me to reconstruct the attributes of elements in the choice set (routes and times of day) for each driver. I then use this reconstructed choice set to estimate the structural model of driving demand outlined in Section 3.

$$\mathbb{1}(h_i = 1 \wedge r_i = 1) = -\alpha T(h_i, r_i) - \gamma_e \underbrace{|h_i + T(h_i, r_i) - h_i^A|_-}_{\text{time early}} - \gamma_l \underbrace{|h_i + T(h_i, r_i) - h_i^A|_+}_{\text{time late}} - \beta \hat{p}(h, r) + \epsilon_{h,r,i} \quad (15)$$

$$p(h_i) = \alpha_0 + \delta_1 post + \delta_2 post * peak + \eta_{h,r,i} \quad (16)$$

Where $\mathbb{1}(h_i = 1 \wedge r_i = 1)$ is an indicator variable that takes a value of 1 if individual i crosses bridge r at time of day h, and zero otherwise. The routes available to a driver are each of the four Bay-Area bridges. Times of day are discretized at 12-minute intervals.

This estimation strategy leverages variation in trip-level attributes that reflect both the 2010 changes an toll prices, as well as differences in the attributes of trips available to drivers across routes or times of day. The identifying variation in price, $p(h_i, r_i)$, comes from the revision to bridge tolls. Peak-hour pricing was imposed on the Bay Bridge in response to high demand for trips connecting Oakland to San Francisco during peak hours. If the high peak-hour demand on this bridge is completely explained by trip attributes — travel time and scheduling costs — then price will be uncorrelated with the structural error term $\epsilon_{h,r,i}$. If, however, this high demand was the result of unobserved shocks to demand unobserved to the researcher, then peak-hour pricing on the bay bridge would create a mechanical correlation between $p(h_i, r_i)$ and $\epsilon_{h,r,i}$. To account for this potential endogeneity, I instrument for price using *post* and *peak* dummies. Said differently, this strategy accounts for the possibility of reverse causality in peak prices. The identifying variation for travel time $T(h_i, r_i)$ comes from both within-day differences in travel time along a given route for each driver, as well as differences in travel times across routes (bridges) conditional on departure time. Variation in travel times in response to the the 2010 change in toll prices is negligible. The schedule cost parameters reflect i) the tradeoff between travel time and late or early arrival induced by variation in travel times throughout the day and ii) the tradeoff between early or late arrival and lower toll prices for peak-hour travelers on the bay bridge.

The estimated parameters of Equation 15 imply a matrix of own and cross-price elasticities between routes and hours of day that I use to solve for second-best cordon prices in San Francisco. Formally, the own and cross-price elasticities from a multinomial logit regression used to estimate this model are:

$$\varepsilon_{\{h^j,r^k\},\{h^l,r^m\}} = \begin{cases} \beta p(h^l, r^m)(1 - s_{\{h^l,r^m\}}), & \text{if } i = l \wedge j = m \\ \beta p(h^l, r^m) s_{\{h^l,r^m\}}, & \text{otherwise} \end{cases} \quad (17)$$

Where $\{h^j, r^k\}$ denotes route r^k taken at time h^j, $s_{\{h^j,r^k\}}$ is the share of total trips taken via route r^k at time h^j, and β and $p(h, r)$ are defined as above. Importantly, logit models exhibit restrictive substitution parameters. Namely, the cross-price elasticities for a given good are constant across all alternatives, implying proportional

substitution following a price increase of any one good. I relax this assumption in my preferred specification — a random coefficients logit regression. This model allows for idiosyncratic pairwise substitution parameters:

$$\varepsilon_{\{h^j,r^k\},\{h^l,r^m\}} = \frac{p(h^l,r^m)}{s_{\{h^j,r^k\}}} \int \beta s_{\{h^j,r^k\}}(\beta) s_{\{h^l,r^m\}}(\beta) f(\beta) d\beta \tag{18}$$

6.2. Bunching Estimator

In this section, I outline how I use notches in the peak-hour tolling on San Francisco's Bay Bridge to recover the scheduling costs of drivers. This alternative empirical approach acts as a check for the results from the logit regressions.

Bunching estimators are used to infer structural parameters from the empirical density of choice variables around kinks or notches in a budget set (Chetty, Friedman, Olsen, and Pistaferri, 2011; Saez, 2010; Kleven and Waseem, 2013). While bunching estimators allow econometricians to identify structural parameters using cross-sectional data, doing so often necessitates strong assumptions regarding the distribution of choice variables under a counterfactual (no-notch) budget set (Blomquist, Newey, Kumar, and Liang, 2021). The panel data in this setting allow me to directly compare the density of trips under notched (peak-hour) and non-notched pricing schemes, thereby circumventing distributional assumptions. Broadly, bunching estimators use changes in the density of choice variables to identify characteristics of a "marginal buncher" — an individual who is indifferent between two positions along a notched/kinked budget set. Before presenting the bunching estimator, it is therefore useful to characterize the marginal bunching individual in this setting.

Consider a group of drivers with homogeneous scheduling elasticities and perfect control over when they cross a bridge. For a marginal buncher, the utility from the lower price is equal to the scheduling costs of adjusting their trip to cross before peak hours. Equation 19 shows this indifference condition in terms of structural parameters:

$$\underbrace{\beta \Delta p}_{\text{Benefit from shifting}} = \underbrace{\gamma_e \Delta h}_{\text{Cost of shifting}} \tag{19}$$

Following the notation from Equation 13, β is the marginal utility of a dollar (normalized to 1), Δp is the change in price at the notch, γ_e is the cost (in dollars/hour) of shifting a trip earlier, and Δh is the number of hours between the price notch and the time of day when the marginal buncher would have crossed the bridge in the absence of a price notch. The scheduling cost, γ_e, can then be written as a function of the size of the price notch, and the distance that the marginal buncher would have to adjust her trip cross the bridge before peak hours:

$$\gamma_e = \frac{\beta \Delta p}{\Delta h} \tag{20}$$

If travel times also differ significantly in the neighborhood of the price notch, this condition becomes:

$$\gamma_e = \frac{\beta \Delta p + \alpha \Delta T}{\Delta h} \tag{21}$$

Where ΔT is the difference between a driver's total travel time if they cross the bridge just before the beginning of peak hours, and a driver's total travel time if they cross the bridge at the time of day when the marginal buncher would have crossed the bridge in the absence of a price notch. The characterization of a marginal buncher is plotted in Figure 6.

Equations 20 and 21 imply that the relevant scheduling cost (either γ_e or γ_l) is inversely proportional to the width of the density trough on the relatively expensive side of the peak-hour price notch. Intuitively, the width

of the density trough reflects how far the marginal buncher moves their trip in response to a price incentive. All else equal, decreasing scheduling costs makes drivers willing to shift their trips further from their ideal travel time for a given level of compensation. A wider the density gap therefore reflects a lower structural scheduling cost.

Because the peak-hour pricing on the Bay Bridge (see Figure 3) creates *notches* rather than *kinks* in the budget sets of drivers, the region immediately adjacent to the price notch is strictly dominated under any scheduling cost. The fact that there is still a positive density of crossings during this dominated period suggests optimization frictions may prevent drivers from perfectly optimizing (Kleven, 2016). In this setting, these 'frictions' may reflect inattentiveness (drivers with automatic tags may not have been aware of the price changes) or the inability to perfectly time bridge crossings due to traffic shocks.[9]

To account for these optimization frictions, as well as heterogeneity in scheduling costs, I use an estimator similar to Kleven and Waseem (2013). I first compare the density of trips in the dominated region before and after the imposition of peak pricing to identify the fraction of individuals with crossing times in the vicinity of the notch who are unresponsive to the price signal. I then estimate the excess trip mass on the relatively inexpensive side of the price notch. Finally, I use these estimates to back out the structural elasticity using the following formula:

$$B = \int_{\gamma_e} \int_{h*}^{h^*+\Delta h} (1-a) f_0(h) dh \simeq (1-a) f_0(h^*) E[\Delta h] \tag{22}$$

Where B is the excess bunching mass on the relatively inexpensive side of the notch, a is the fraction of drivers in the strictly dominated region, and $f_0(h)$ is the counterfactual (no-notch) density of vehicle crossings as a function of the time of day, h. $E[\Delta h]$ is the average adjustment among drivers who bunch at the price notch. Solving Equation 22 for Δ_h and plugging into Equation 21 yields the bunching estimator:

$$\gamma_e = \frac{\beta \Delta p}{B/((1-a) f_0(h^*))} \tag{23}$$

Relaxing the assumption that travel times are relatively flat around the notch point is straightforward, but necessitates the value of travel time:

$$\gamma_e = \frac{\beta \Delta p + \alpha \Delta T}{B/((1-a) f_0(h^*))} \tag{24}$$

[9]Peak hour prices are displayed prominently above the toll plaza on the Bay Bridge, suggesting that is it more likely that "frictions" reflect traffic shocks rather than salience.

Figure 6 — Bunching Theory

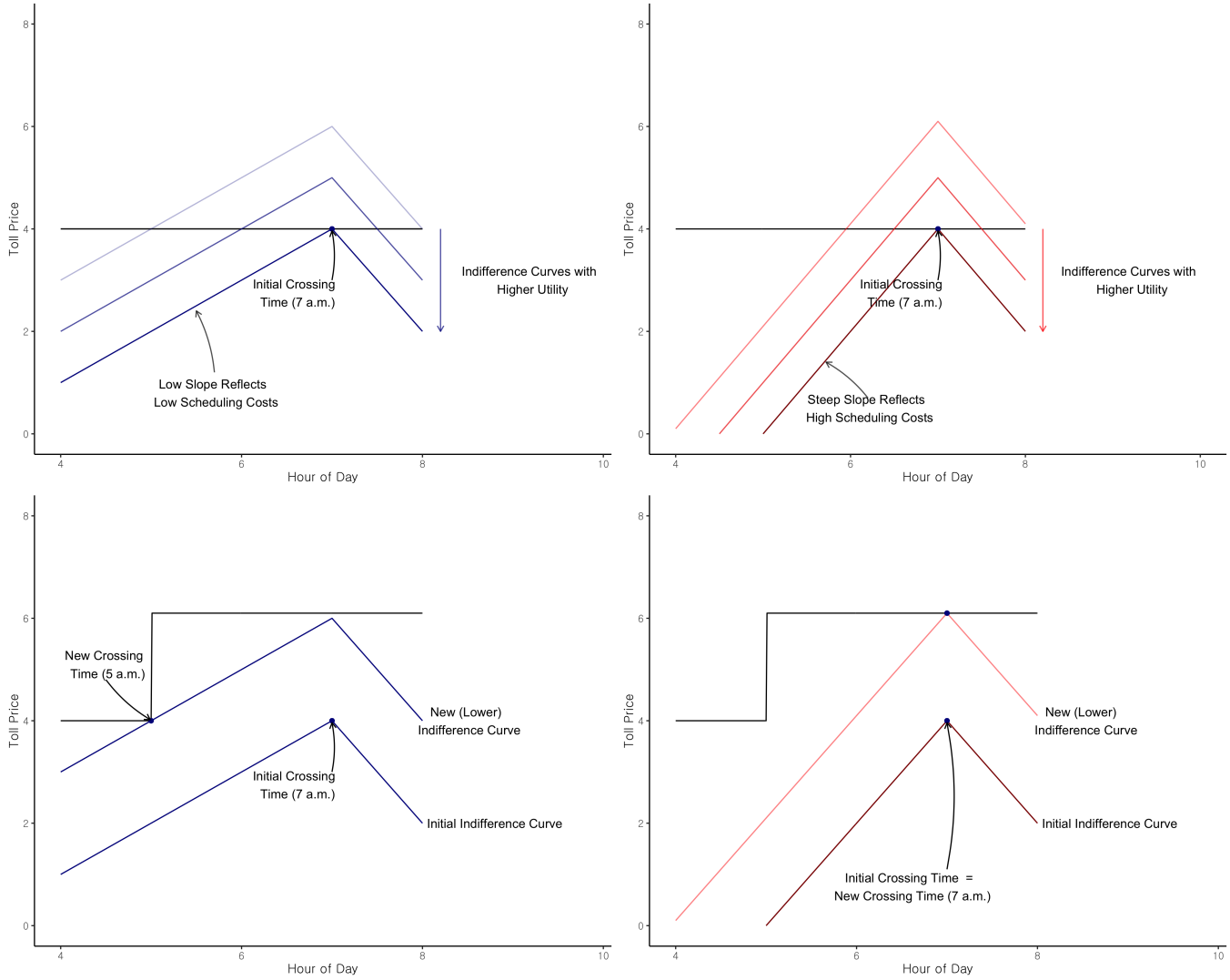


Figure 6: This figure illustrates the determinants of bunching behavior in peak-hour toll schemes based on scheduling costs, as predicted by the structural model outlined in Section 3. For expositional ease, this figure plots the case where travel times are constant throughout the day. In the right two panes, I plot indifference curves (red) of drivers in the case of *high* scheduling costs. The triangular shape of the indifference costs reflects the fact that the further a trip is from a given driver's ideal crossing time, the higher the compensation (via a lower toll price) required to maintain any given level of driver utility. In the left two panes, I plot indifference curves for in the case of *low* scheduling costs. All else equal, when scheduling costs are lower, drivers are more willing to adjust their travel times in response to peak pricing, implying a larger mass of trips around price notches.

7. Results

In this section, I present estimates of the parameters in the structural model of travel demand (13). Broadly, estimates using the FasTrack data suggest that drivers in this setting exhibit a value of travel time in line with government values and existing academic studies ($14 per hour to $23 per hour), and are more willing to shift trips earlier than they are willing to shift trips later.

7.1. Logit Regression Results

Table 1 presents the results from Equation 15, estimated via multinomial logistic regression. The point estimates from this regression suggest that the average driver is indifferent between saving $23 and saving an hour of travel time; they they are indifferent between saving $12 arriving an hour early, and they are indifferent between saving $19 arriving an hour late.

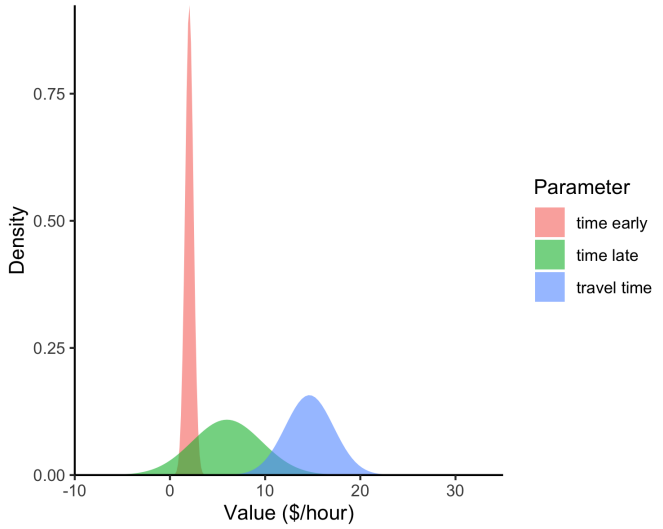
Figure 7 shows results of a mixed logit. Allowing for heterogeneity in structural parameters produces results that are qualitatively similar to the results in Table 1. In Table 3, I allow price responsiveness to vary with road user's idiosyncratic externalities. To do so, I break FasTrak devices into quartiles based on the average estimated externality (both pollution and congestion) of each device's trips. This table suggests that drivers who travel in clean cars or travel on uncongested routes tend to be less elastic than to price than are those who travel in dirty cars or on congested routes.

Table 1 — Structural Model of Driving Demand

variable	**coef**	**se**	**t**	**p**
travel time	23.292	0.232	−6.193	0.000
time early	12.657	0.035	−22.481	0.000
time late	19.764	0.059	−20.648	0.000
price	1.000	0.054	−1.136	0.256
(Intercept)	67.099	0.295	−14.059	0.000

Table 1: Results from Equation 15, a discrete choice model where drivers choose over routes and times of day, estimated using a 10% sample of the FasTrak tolling microdata described in section 5. The dependent variable is whether an individual i elects to take a trip on route r at time of day h. *Travel time* is the travel time (in hours) that driver i would incur by traveling via route r at time h. *Time early* is the number of hours that that driver i would arrive before their ideal arrival time if they were to travel via route r at hour h. *Price* is the toll that driver i would incur by traveling via route r at hour h. To account for possible reverse causality in peak-hour pricing, *post* and *peak* dummies act as instruments for *price*. As the coefficient on price is normalized to 1, the coefficients can be interpreted as dollars per hour.

FIGURE 7 and Table 2 — MIXED LOGIT RESULTS



variable	**mean**	**sd**
time early	7.368	0.683
time late	23.816	2.692
travel time	44.926	2.081
price	1.000	0.175

Figure 7: This figure displays the mean and standard deviation from a random coefficients ("mixed") logit regression used to estimate the parameters in Equation 13. The data used in this regression are a 10% sample of the FasTrak tolling microdata described in section 5. The dependent variable is whether an individual i elects to take a trip on route r at time of day h. *Travel time* is the travel time (in hours) that driver i would incur by traveling via route r at time h. *Time early* is the number of hours that that driver i would arrive before their ideal arrival time if they were to travel via route r at hour h. *Price* is the toll that driver i would incur by traveling via route r at hour h. To account for possible reverse causality in peak-hour pricing, *post* and *peak* dummies act as instruments for *price*. The mean and standard deviation of all time-related variables have been normalized relative to the coefficient on the *price* variable.

Table 3 — MIXED LOGIT WITH PRICE ELASTICITIES BY EXTERNALITY QUARTILE

variable	**mean**	**sd**
time early	−1.247	0.488
time late	−3.409	2.078
travel time	−5.371	1.570
price (first externality quartile)	−0.016	NA
price (second externality quartile)	−0.148	NA
price (third externality quartile)	−0.285	NA
price (fourth externality quartile)	−0.126	NA

Table 3: Results from Equation 15, a discrete choice model where drivers choose over routes and times of day. The dependent variable is whether an individual i elects to take a trip on route r at time of day h. *Travel time* is the travel time (in hours) that driver i would incur by traveling via route r at time h. *Time early* is the number of hours that that driver i would arrive before their ideal arrival time if they were to travel via route r at hour h. *Price* is the toll that driver i would incur by traveling via route r at hour h. I interact *price* with *externality quartile*, a categorical variable defined at the individual level that indicates the average intensity of externalities (both pollution and congestion) for trips taken by each device in the FasTrak dataset.

7.2. Bunching Estimator Results

Applying a bunching estimator to notches in the pricing schedule on the Bay Bridge, I recover scheduling cost parameters (γ_e and γ_l in Equation 13) that accord with the estimates from the logit regressions. Drivers are generally more willing to shift trips earlier than later, with point estimates of scheduling costs that range from \$6 to \$15 per hour.

Figure 8 plots the frequency of car trips by time of day before versus after the imposition of peak hour pricing; figure 9 plots the difference between the density of trips by time of day before vs. after the imposition of peak-hour pricing. Both figures show distinct spikes in the density of trips immediately outside of the peak-hour pricing window and accompanying gaps in the density of trips taken just inside the peak-hour pricing window. Qualitatively, the bunches appear to be most pronounced during the early morning (5 a.m.) and evening (7 p.m.) price notches. Intuitively, this suggests that it is less costly to arrive early or leave late from work than it is to arrive late or leave early. During morning commute hours, the marginal buncher is roughly indifferent between saving \$6 being an hour early, and indifferent between saving \$15 and being an hour late. During evening commute hours, the marginal buncher is roughly indifferent between saving \$9 being an hour early, and indifferent between saving \$13 and being an hour late.

Figure 8 — Bunching at Price Notches

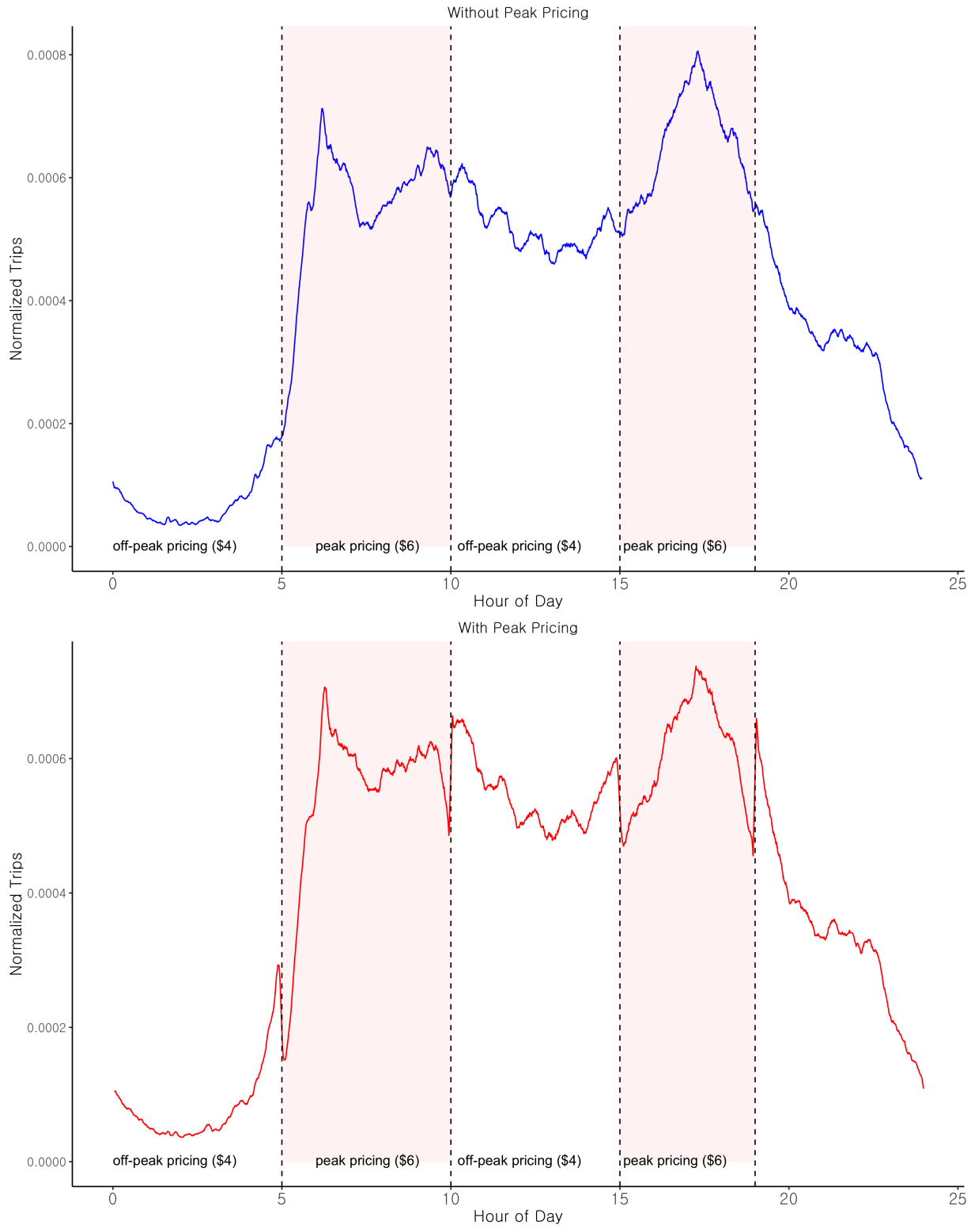


Figure 8: This figure plots the density of passenger vehicle trips crossing the Bay Bridge in the 6 months before (blue) and 6 months after (red) the imposition of peak hour pricing on July 1st, 2010. This plot excludes trips that use the carpool lane, as well as eligible electric vehicles, each of which faced a different pricing scheme.

Figure 9 — Post vs. Pre Bunching

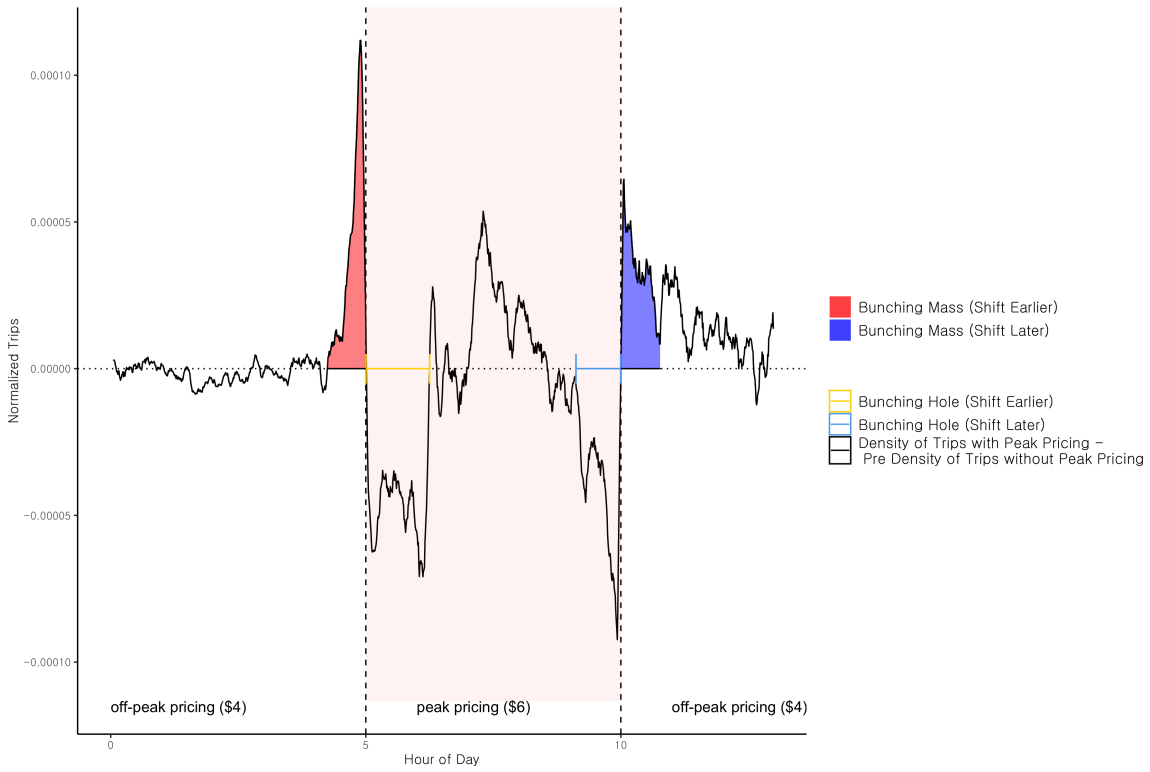


Figure 9: This figure plots the difference in density of trips in the 6 months before vs the 6 months after the imposition of peak-hour pricing on the Bay Bridge on July 1st, 2010. The horizontal bars plot the width of the bunching hole adjacent to each price notch. Equation 20 suggests that the width of this density hole is inversely proportional to the structural scheduling costs. The shaded regions represent the excess mass of trips taken just outside the peak pricing window. I use the mass in these regions to estimate scheduling costs using equation 22.

Table 4 — Estimating Scheduling Costs via Bunching

Parameter	**Estimate**
Schedule cost early, morning	6.19461
Schedule cost early, evening	9.74395
Schedule cost late, morning	15.49765
Schedule cost late, evening	13.75856

Table 4: Rows of this table show estimates of the costs of adjusting scheduling car trips earlier or later (γ_e and γ_l in equation 13), estimated using the amount of bunching at price notches on San Francisco's Bay Bridge (equation 22). The estimates in this table reflect both scheduling frictions, as well as any time savings that also resulted from drivers adjusting their trips to fall just outside of peak hours (assuming a $20 value of travel time).

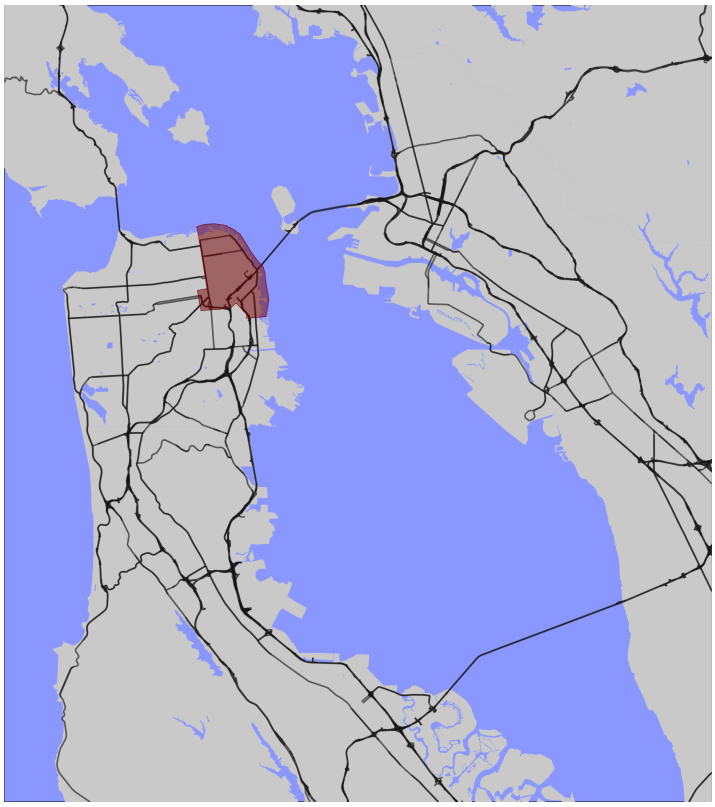
8. Second-Best Optimal Cordon Prices

In this section I use the discrete choice model estimated in Section 7 (Table 3) to calculate optimal cordon zones. I first demonstrate this framework using San Francisco's proposed cordon, and then consider cordon zones in Los Angeles and New York. At a high level, calculating optimal cordon prices in any city takes four steps: First, use travel survey data (e.g., the NHTS) to identify a representative sample of trips that pass through a city's proposed cordon. Second, assign externalities to those trips using information about the vehicle driven in each trip and the traffic density along the trip (this process is similar to the process describes in section 5). Third, apply the model estimated in Section 7 to calculate substitution parameters. And fourth, apply the optimal tax formula outlined in Section 2 to the ingredients from steps 1-3.

8.1. San Francsico's Proposed Cordon zone

The San Francisco County Transportation Authority (SFCTA) intends to pilot a downtown congestion pricing program in the next 3-5 years, with the goal of implementing cordon pricing by the end of the decade (San Francisco County Traffic Authority, 2021). Figure 10 shows a map of the proposed cordon zone, along with the proposed tolling schedule.

FIGURE FIGURE 10 — SAN FRANCISCO'S PROPOSED CONGESTION PRICING SCHEME



Income Group	High Income	Middle Income	Low Income	Very Low Income
Peak Hours (6-9am; 3:30-6:30 pm)	$6.50	$4.33	$2.17	Free
Off-Peak Hours	Free	Free	Free	Free

Figure 10: San Francisco's proposed pricing scheme as of September 1st, 2021. Trips that enter the cordon would be charged during peak hours according to the income of the registered vehicle. An individual's *Income Group* depends both on income and family size. For single individuals, for example, the annual income thresholds for high, middle, low, and very low income are 150k, 116k, 66k and 46k, respectively.

As outlined in Section 2, calculating the second-best optimal prices in this setting requires information about the marginal damages of trips that cross through a cordon zone, as well as information about the elasticity and substitutability of these trips. In this section, I use trip-level data from the 2017 NHTS California Add-On to construct a dataset of trips that cross San Francisco's proposed cordon zone. I then use the results from Section 7 to predict how substitutable these trips are in time and space, allowing me to identify the optimal pricing scheme for San Francisco's cordon zone.

8.2. Trips in the San Francisco Area

I use the 2017 NHTS California Add-On to build a representative dataset of trips that cross San Francisco's proposed cordon zone.[10] Each *trip* in this dataset consists of a start location (zip code or Census Block), an end

[10]The FasTrak microdata are ill-suited for this task because many trips that cross the cordon do not cross any bridge (e.g., trips with two termini within the city of San Francisco).

location (zip code or Census Block), information about the vehicle that took the trip (make, vintage, fuel type), and the time of day that the trip was taken. I determine whether or not a trip passes through the cordon using the HERE Technology's *Routes* API. This dataset has 1,891 routes that cross the cordon zone during weekdays between the hours of 4 a.m. and 10 p.m., which I plot in Figure 11.

To predict substitution in time and space, I construct a set of alternatives for each trip. For every NHTS trip, I construct alternative departure times at 15-minute intervals throughout the day. Using HERE Technology's *Routes* API, I can assign travel times to each of these alternative trips. For trips with termini that lie outside of the cordon zone (i.e., trips that only pass through the cordon zone en route to their destination) I identify the most direct detour that circumvents the cordon zone — if both the start and end coordinates of the trip are outside of the cordon, I use HERE's Routes API to find the most direct route that does not cross the cordon. I then calculate travel times for this non-cordon route for each 15-minute interval throughout an average traffic day.

The result of this data collation is a set of trip endpoints for the San Francisco area, where drivers can chose over $route \in \{\text{cordon, non-cordon}\}$ and $time\ of\ day \in \{4.0, 4.1, \ldots 22.0\}$ based on the same attributes from Section 7: travel time, time early, time late, and toll price.

Figure 11 — NHTS Routes and Cordon Detours

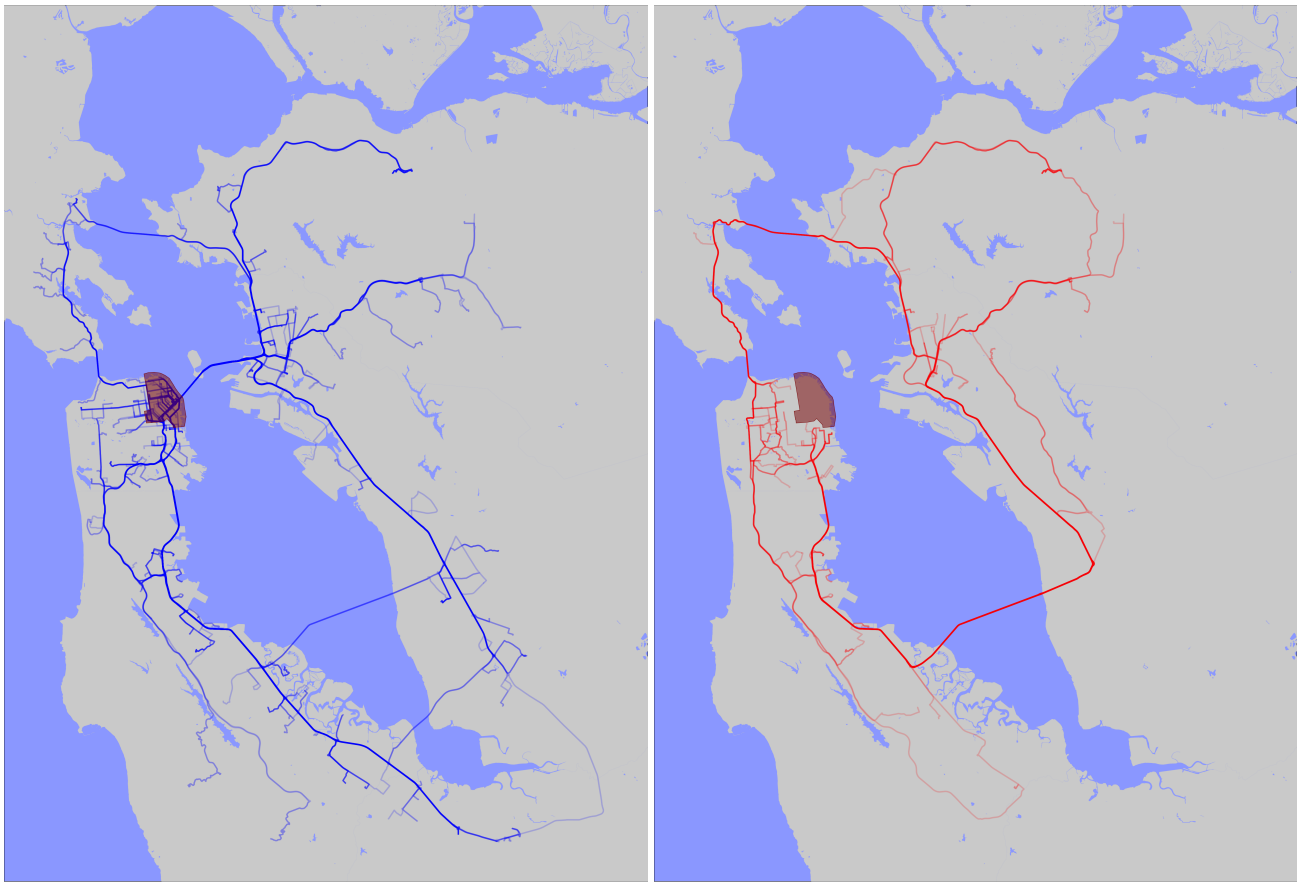


Figure 11: This figure plots vehicle trips from the 2017 National Household Transportation Survey (NHTS) California Add-On. The left pane plots routes that cross San Francisco's proposed congestion zone according to suggested routes generated with the HERE Technology's *Routes* API. The right pane plots detour routes for trips where it is possible to circumvent the congestion zone (i.e., trips with both start and end points that are outside of the cordon). Each driver's choice set consists of a cordon route (the left pane) for every 6-minute time of day interval, as well a non-cordon route (the right pane) if such a detour exists, for every 6-minute time of day interval.

8.3. Trip-Level Externalities

For each trip described above (both trips listed in the NHTS and alternative trips in space/time), I assign traffic and pollution externalities in a manner similar to the process described in Section 5.

As shown in Figure 5, emissions vary by vehicle attributes as well as travel speed. The NHTS includes information about each vehicle, including the vehicle vintage, make, and fuel type (gasoline, diesel, EV, or hybrid). The travel time and distance information for each trip returned by the HERE traffic API allow me to assign an average speed to each trip. I then merge emissions factors onto each trip based on vehicle vintage, fuel type, and travel speed, using data from California's EMFAC database. I plot the emissions externalities for the 1,891 NHTS trips that cross the proposed Cordon in Figure 12.

To assign congestion externalities to trips, I use estimates from Yang, Purevjav, and Li (2020), who show that the marginal external (travel time) cost of traffic is convex in traffic density. To determine the density along a given route in my dataset, I use a comprehensive network of traffic sensors on roadways in the Bay Area. For each trip, I identify sensors along the trip's route. I then assign a marginal external congestion cost (in dollars

per mile) to this point based on the average traffic density at that sensor at the time of day associated with the trip. A trip's total congestion externality is then the average of the external congestion costs (in dollars per mile) along the route times the length of the trip. I plot the trip-level externalities for the 1,891 NHTS trips that cross the proposed Cordon in day in Figure 13.

Figure 12 — Pollution Costs for Trips Crossing San Francisco's Cordon

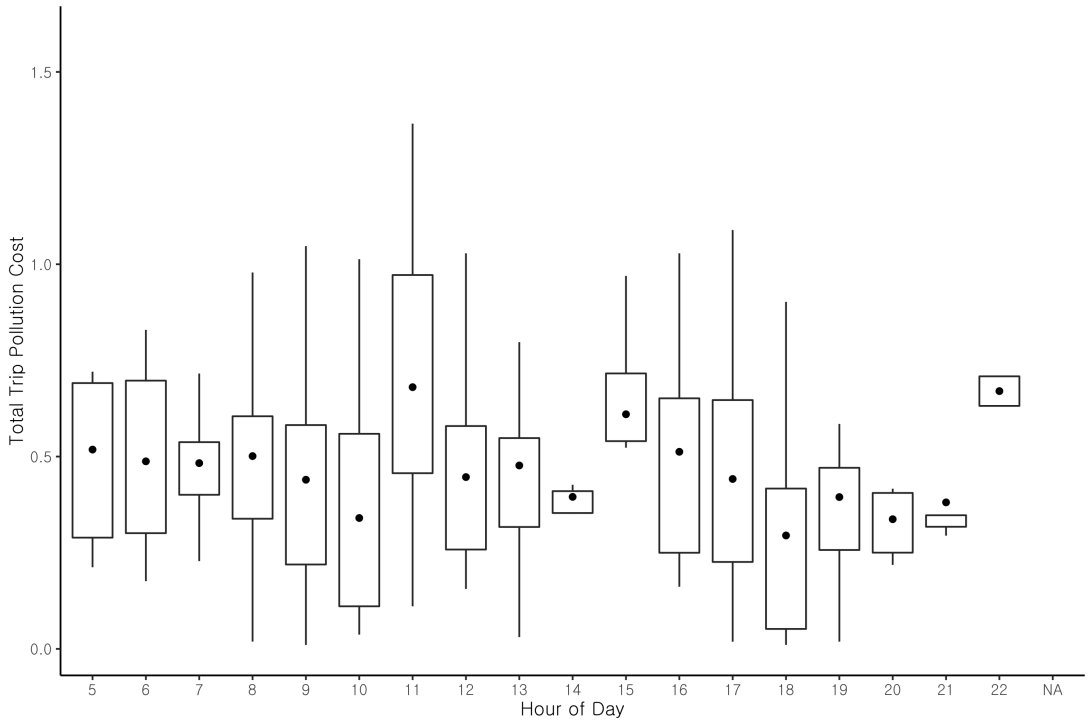


Figure 12: This Figure plots pollution costs by hour for trips in the 2017 National Household Transportation Survey (NHTS) that pass through San Francisco's proposed cordon zone. Trip routes reflect the suggested routes from HERE Technology's Routes API. Pollution emissions were calculated by merging emissions factors from California Air Resources Board's EMFAC database to trips based on vehicle fuel type, vehicle age, and average trip travel speed. I convert emissions to externalities using EPA social costs for global pollutants and EAISUR costs for local pollutant emissions in San Francisco. All values are in 2020 dollars.

Figure 13 — Congestion Costs for Trips Crossing San Francisco's Cordon

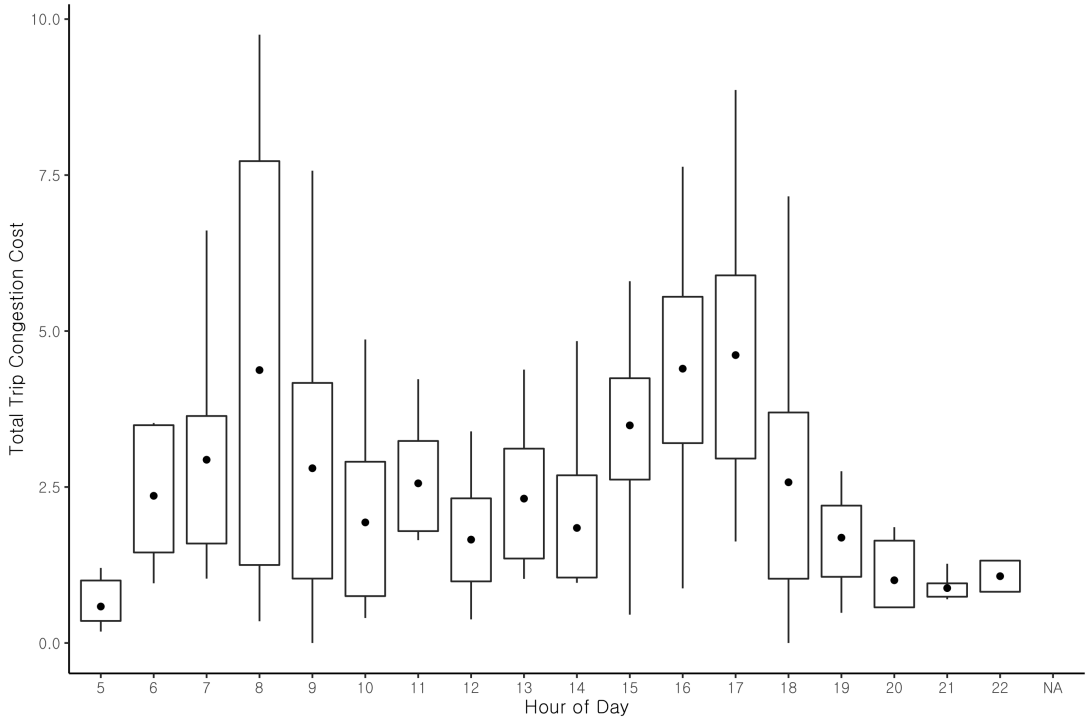


Figure 13: This Figure plots congestion costs by hour for trips in the 2017 National Household Transportation Survey (NHTS) that pass through San Francisco's proposed cordon zone. Trip routes reflect the suggested routes from HERE Technology's *Routes* API. Congestion costs were calculated by identifying traffic sensors along a given route and assigning per-mile congestion costs to each sensor using estimates of the density-congestion relationship from Yang, Purevjav, and Li (2020). All costs are in 2020 dollars.

8.4. Substitution Parameters

The last set of parameters necessary for calculating optimal cordon prices are the parameters that govern how substitutable trips are in time and space. Specifically, calculating optimal prices using 10 requires *leakage shares* between trips: $\frac{dh_k}{dp_j} / \frac{dh_j}{dp_j}$. Recall that if j and k are trips (defined as a specific route $\in$ {cordon, non-cordon} at a specific hour of day $\in$ {4.1, 4.2, ..., 22.0}) the leakage share between trip k and trip j represents the share of the reduction in usage of trip k that shifts to trip j as a result in the increase of the price of taking trip j. For a concrete example, imagine that a one dollar increase in the price of driving through a cordon zone between the hours of 8 a.m. and 9 a.m. reduces trips by 10%, with 6% of all trips shifting one hour earlier (call these trips y) and 4% of trips shifting to routes circumvent the cordon (call these trips z). The leakage shares for this example for trips x and y are $\frac{dh_y}{dp_x} / \frac{dh_x}{dp_x} = 0.6$ and $\frac{dh_z}{dp_x} / \frac{dh_x}{dp_x} = 0.4$, respectively.

The leakage shares are implied directly from parameters of the mixed logit regression[11]. Formally, for any two trips $\{h^l, r^m\}$ and $\{h^j, r^k\}$, where h is the hour of day for a given trip and r is an indicator for whether or not the trip crosses San Francisco's cordon, the leakage share between these two trips is:

$$\frac{\frac{\partial\{h^j,r^k\}}{\partial p\{h^l,r^m\}}}{\frac{\partial\{h^l,r^m\}}{\partial p\{h^l,r^m\}}} = \int \beta s_{\{h^j,r^k\}}(\beta) s_{\{h^l,r^m\}}(\beta) f(\beta) d\beta \tag{25}$$

where β is the joint distribution of random coefficients in the mixed logit model, and $s_{\{h^j,r^k\}}(\beta)$ is the share predicted trips that take route k at time j under the random coefficient vector β.

8.5. Optimal Prices

Figure 14 plots three prices necessary for understanding optimal cordon prices. The blue (solid) line plots the average externalities for trips that pass through San Francisco's cordon zone by hour of day, estimated using the process detailed above. The green (dotted) line shows these externalities re-weighted as per Diamond (1973) to account for the correlation between the price-responsiveness of trips and idiosyncratic trip level externalities, as reported in Table 3. Finally, the red line plots the second-best optimal price for San Francisco's proposed cordon when tolling is restricted to morning and evening peak hours (6-10am and 3-7pm, respectively). The second-best optimal price is calculated using equation 10, and takes into account both the correlation between externalities and elasticities, as well as the substitution to unpriced alternatives in time and space.

[11] Note the distinction between this formula, which recovers a cross-price *derivative*, and the canonical formula for a mixed logit cross-price *elasticity* (e.g., page 144 of Train (2009))

Figure 14 — Second-Best Optimal Road Prices

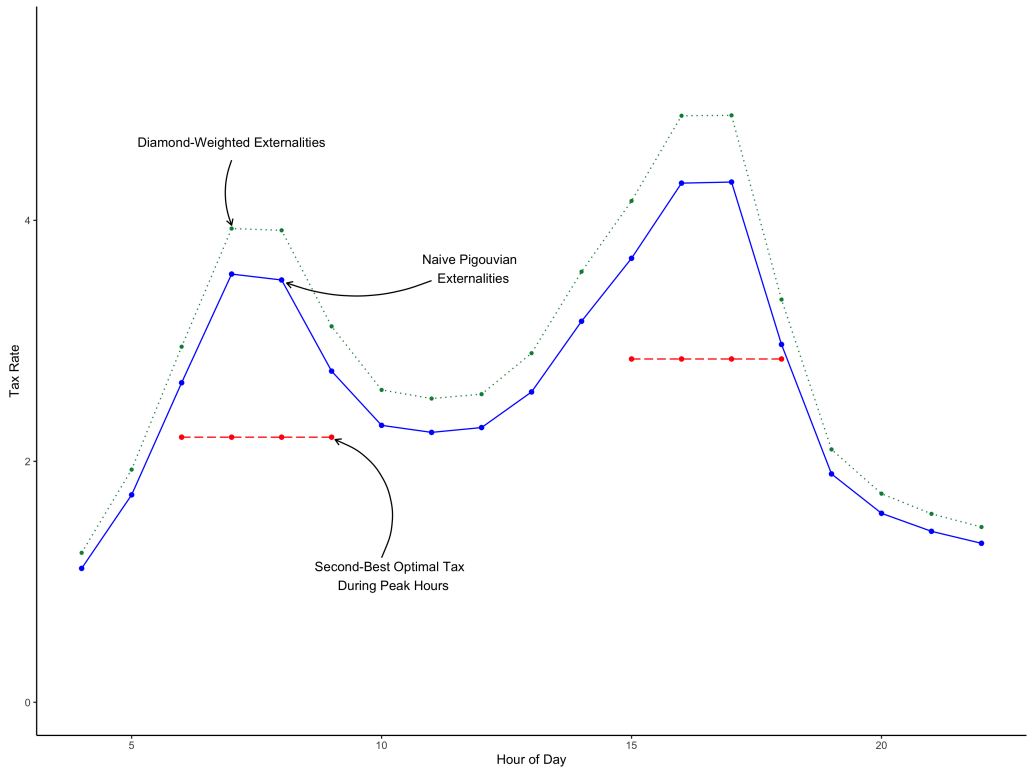


Figure 14: This figure plots three prices relevant for understanding optimal second-best cordon prices. The blue (solid) line plots the average externality (pollution and congestion) for trips that cordon San Francisco's cordon by hour of day, estimated using data from the 2017 NHTS. The green (dotted) line plots externalities re-weighted to account for the correlation between trip-level externalities and trip-level elasticities, as per Diamond (1973). The red line plots the second-best optimal price for San Francisco's proposed cordon when tolling is restricted to morning and evening peak hours (6-10am and 3-7pm, respectively). The second-best optimal price is calculated using equation 10, and takes into account both the correlation between externalities and elasticities ("Diamond externalities"), as well as the substitution to unpriced alternatives in time and space.

8.6. The Impact of Pricing on Congestion, Emissions, and Welfare

After augmenting the NHTS data with information about externalities and the attributes of trips with alternative routes and travel times, I simulate travel decisions in three scenarios.

1. **No congestion pricing**. This is the status quo; the only charges that trips may face are the existing Bay Area bridge tolls, set to 2020 levels.

2. **First-best (Pigouvian) pricing**. Every trip a driver could choose would be priced according to its social damages, which include both congestion and pollution externalities.

3. **Second-best optimal peak-hour cordon prices**. These prices are calculated using equation 10. Trips that pass through the cordon area are charged $2.10 during morning peak hours, and $2.90 during evening peak hours (see Figure 14).

I plot outcomes from these simulations in Figures 15 through 17, and summarize the results in Table 5. Two themes emerge: first, on all three outcome measures — trips, congestion externalities, and pollution externalities — second-best optimal peak-hour pricing more closely resembles the status quo than the first-best policy. Second, there are distinct bunches in trips, congestion, and pollution just outside peak-hour pricing periods.

Figure 15 — Simulated Travel Choices Under Road Pricing

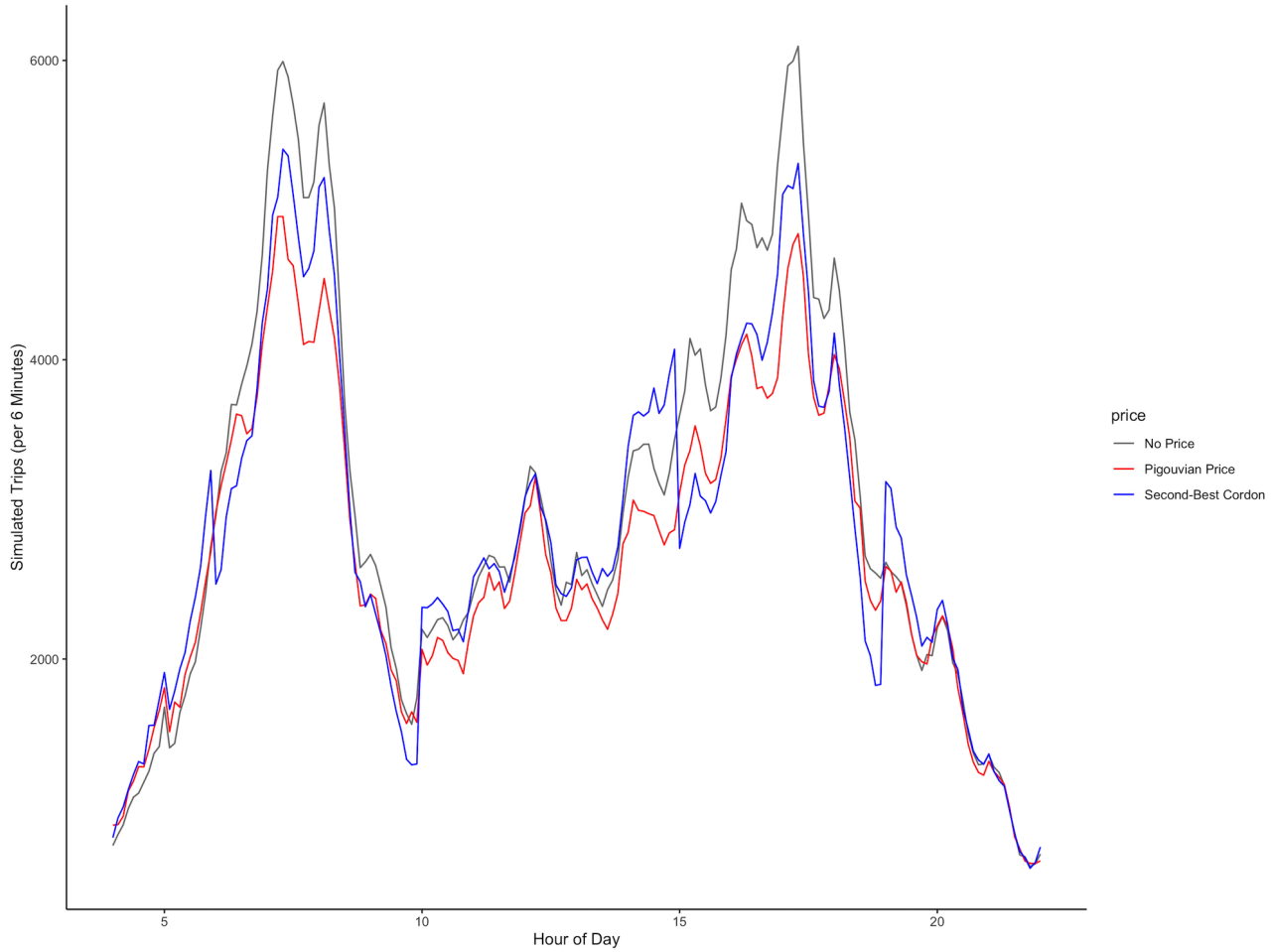


Figure 15: In this figure I plot the number of trips that pass through or near the cordon under three simulations using the mixed logit model estimated in Table 3 of Section 7 together with the NHTS trip dataset described in Section 8. In each scenario, I predict 600,000 choices — roughly daily total of vehicle trips that pass through San Francisco's proposed cordon San Francisco County Traffic Authority (2021). The grey line plots predicted trips by time of day without any pricing (the status quo). The blue line plots trips under the (infeasible) first-best scenario where every trip a driver could choose would be priced according to it's marginal pollution and congestion externalities. The red line plots trips under the second-best optimal peak-hour cordon price from Figure 14. Note that all lines include both trips that cross through the cordon, and "detour" trips that circumvent the cordon.

FIGURE 16 — SIMULATED CONGESTION EXTERNALITIES CHOICES UNDER ROAD PRICING

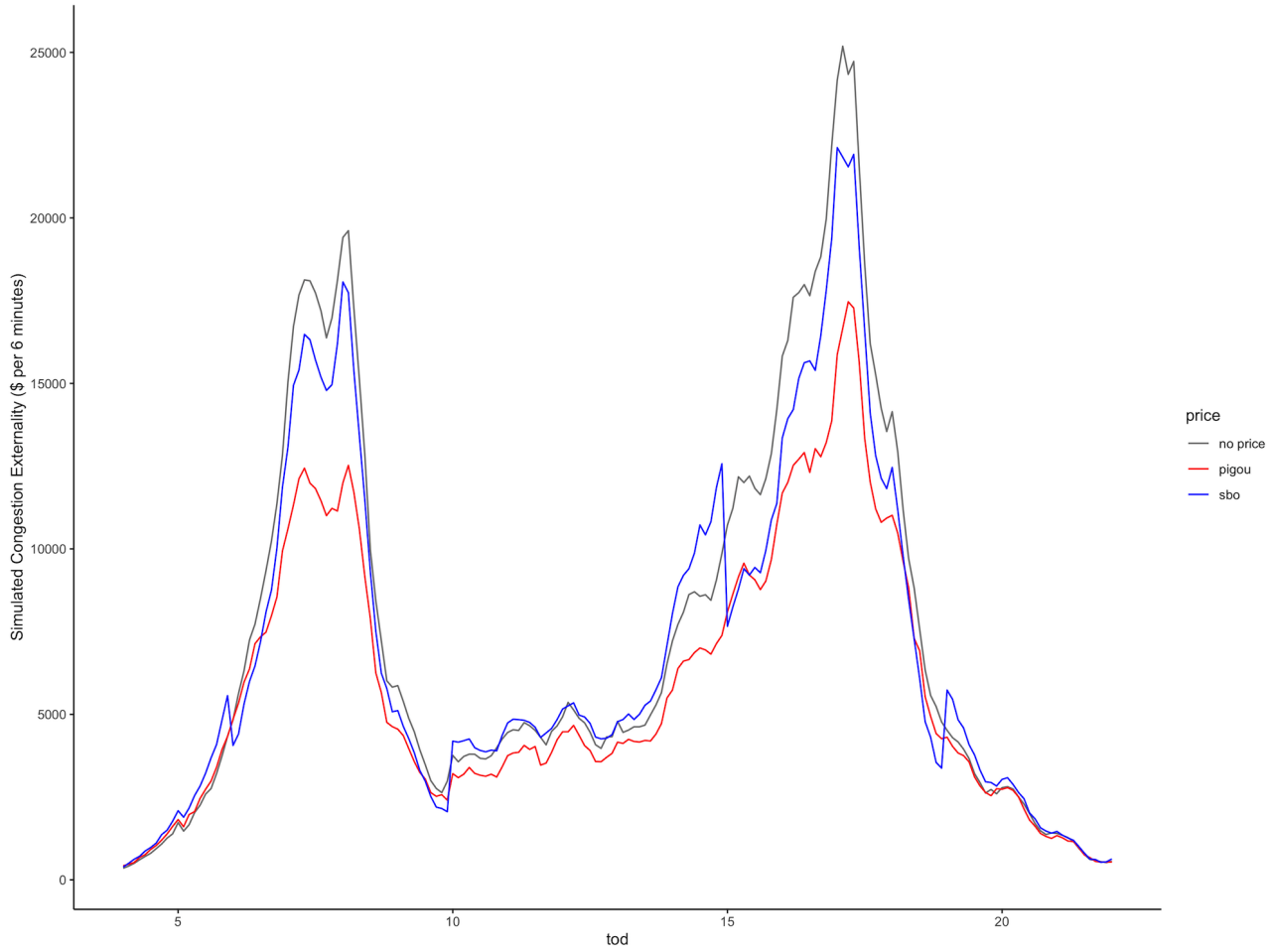


Figure 16: In this figure I plot the total congestion externalities under three simulations using the mixed logit model estimated in Table 3 of Section 7 together with the NHTS trip dataset described in Section 8. In each scenario, I predict 600,000 choices — roughly daily total of vehicle trips that pass through San Francisco's proposed cordon San Francisco County Traffic Authority (2021). The grey line plots the sum of congestion externalities by time of day without any pricing (the status quo). The blue line plots congestion under the (infeasible) first-best scenario where every trip a driver could choose would be priced according to it's marginal pollution and congestion externalities. The red line plots sum of congestion externalities under the second-best optimal peak-hour cordon price from Figure 14. Note that all lines include congestion from trips that cross through the cordon, as well as "detour" trips that circumvent the cordon.

Figure 17 — Simulated Pollution Externalities Choices Under Road Pricing

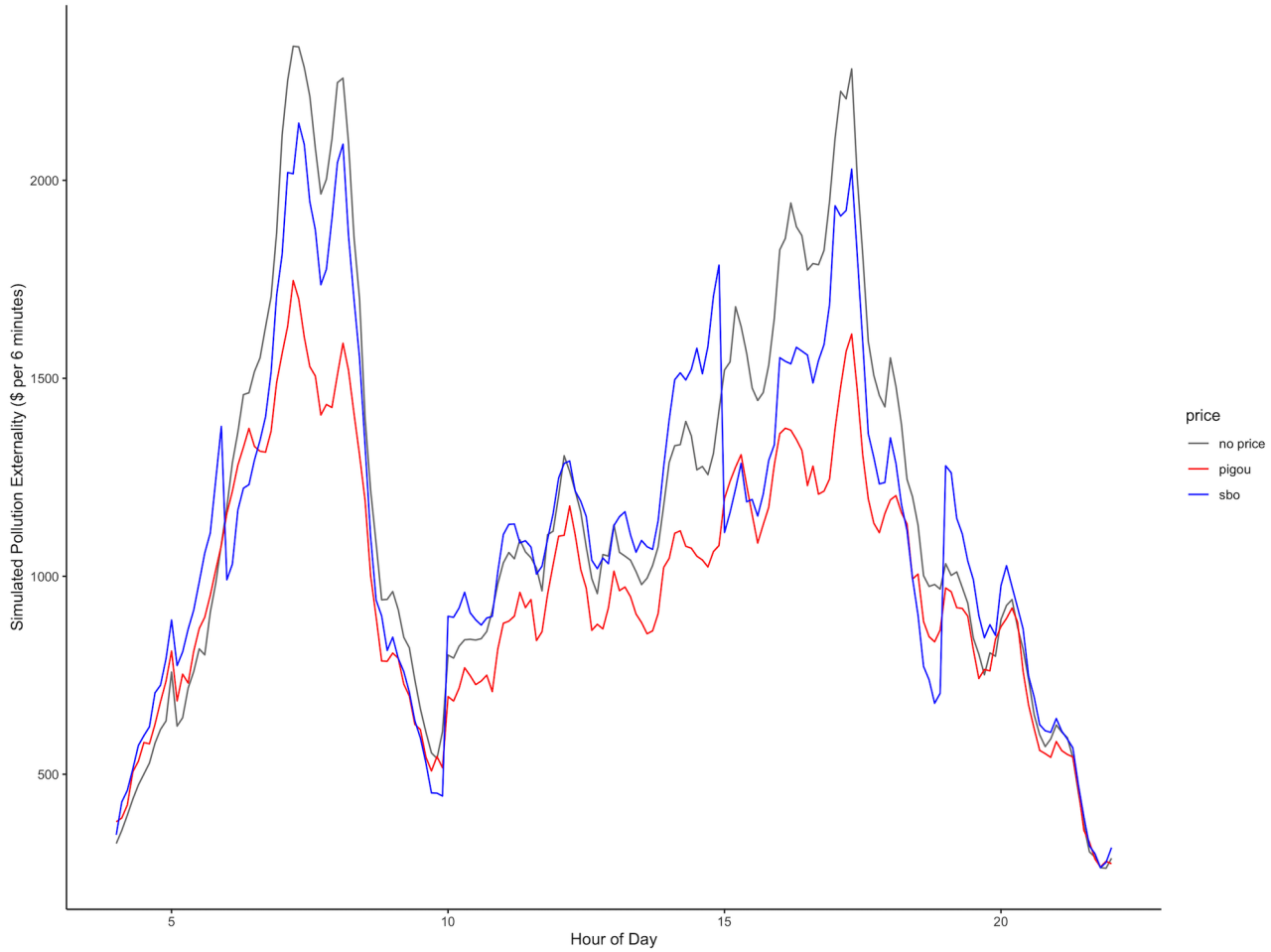


Figure 17: In this figure I plot the total pollution externalities under simulations using the mixed logit model estimated in Table 3 of Section 7 together with the NHTS trip dataset described in Section 8. In each scenario, I predict 600,000 choices — roughly daily total of vehicle trips that pass through San Francisco's proposed cordon San Francisco County Traffic Authority (2021). The grey line plots the sum of pollution externalities by time of day without any pricing (the status quo). The blue line plots pollution externalities under the (infeasible) first-best scenario where every trip a driver could choose would be priced according to it's marginal pollution and congestion externalities. The red line plots sum of pollution externalities under the second-best optimal peak-hour cordon price from Figure 14. Note that all lines include pollution from trips that cross through the cordon, as well as "detour" trips that circumvent the cordon.

Table 5 — Congestion, Pollution, and Welfare Effects of San Francisco's Cordon Zone

Outcome	Performance Relative to First-Best (%)
Reduction in Total Externalities	30.276
Reduction in Congestion	31.043
Reduction in Pollution	23.699
Welfare Gain	28.840

Table 5: This table compares the second-best optimal cordon pricing scheme in San Francisco to an (infeasible) first-best policy where vehicles are charged according to the total marginal damages associated with Each trip. The four outcomes of interest are total externalities (pollution and congestion), congestion alone, pollution alone, and total welfare (the utility of drivers, in dollars, less total externalities). The figures in this table reflect 600,000 simulated choices (roughly the number of weekday trips that pass through San Francisco's proposed cordon) using the mixed logit model shows in Table 3.

8.7. Cordon Pricing in New York and Los Angeles

City governments in New York and Los Angeles currently considering cordon pricing zones (mapped in Figure 18). In this section, I calculate optimal cordon prices for each of these cities, and evaluate the performance of the second-best optimal cordon pricing scheme relative to a policy that prices every trip at social marginal damages.

Figure 18 — Proposed Cordons in New York and Los Angeles

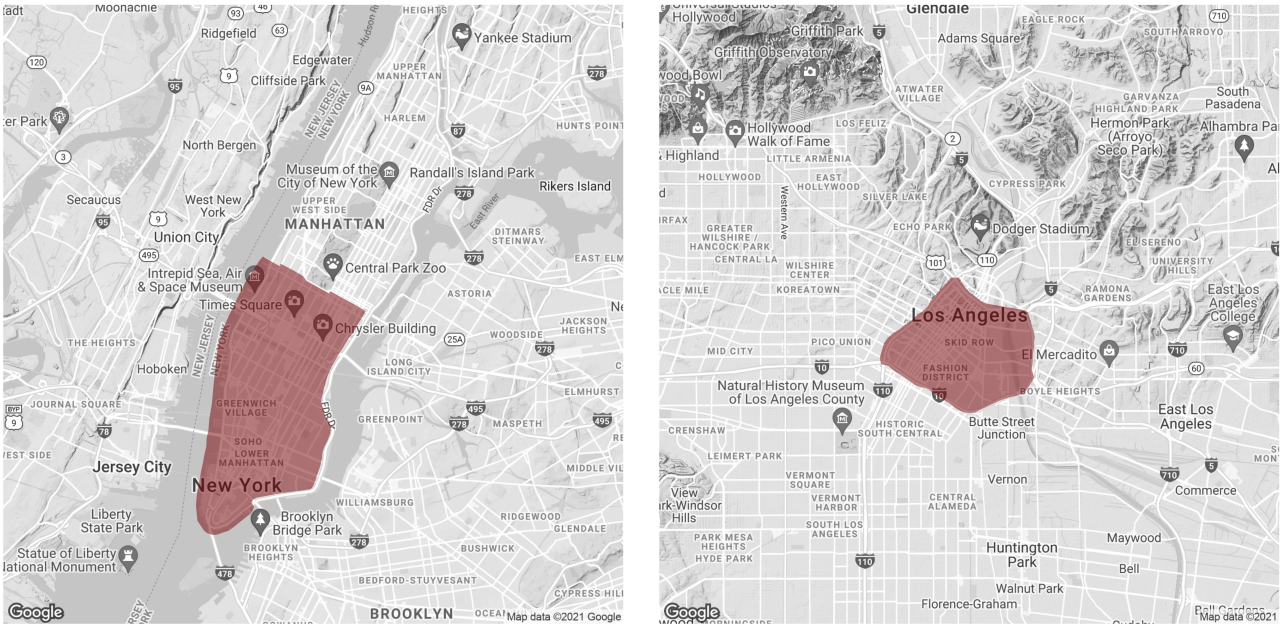


Figure 18: Proposed cordon pricing schemes in New York and Los Angeles. All proposals are as of August, 2021. The New York congestion map is courtesy of the *Regional Plan Association*; The Los Angeles map is courtesy of the *LA Metro*.

As outlined in Section 2, calculating the second-best optimal cordon prices requires information about the marginal damages of trips that cross through a cordon zone, as well as information about the elasticity and

substitutability of these trips. For each of the above cities, I follow the same general template as in San Francisco (see Sections 8.2 through 8.4): First, I use survey data[12] and Here Technology's *Routes* API to identify trips that would use each city's cordon. Second, I use vehicle attributes and travel speed to assign pollution externalities, and use traffic density data[13] from city roads to assign congestion externalities to those trips. Third, calculate substitution parameters between those trips.

Ideally, there would be a natural experiment in each city that would allow for the estimation of city-specific travel demand primitives (scheduling costs, γ_e and γ_l, and the value of travel time α) that are used to calculate substitution parameters, as well city-specific correlations between externalities and price responsiveness (Diamond weights). Absent such experiments, I calculate optimal cordon prices under two scenarios. In the first scenario (Table 7), I use the travel demand parameters and Diamond weights estimated in San Francisco (see Table 3). In the second scenario (Table **??**), I use questions from the 2017 NHTS to adjust model primitives and externality-elasticity correlations to match survey responses in each city. Specifically, the NHTS asks respondents to report their schedule flexibility (Yes/No) as well as their responsiveness to gasoline demand (Scale of 1 to 5). I use the former question as a proxy for scheduling elasticity, and the latter as a proxy for price responsiveness. For further anecdotal evidence of the external validity of the model estimated in San Francisco, see Appendix G.

Table 6 — Congestion, Pollution, and Welfare Effects of Peak-Hour Cordon Pricing

	Value ($)		
Period	San Francisco	Los Angeles	New York
Second-Best Price, AM Peak (6-10)	2.201	3.298	7.294
Second-Best Price, PM Peak (3-7)	2.850	4.533	7.919
Average Marginal Damages, AM Peak (6-10)	3.115	4.877	8.186
Average Marginal Damages, PM Peak (3-7)	3.821	5.724	12.635

Table 6: This table compares second-best optimal peak hour prices for the proposed cordons in San Francisco, Los Angeles, and New York to the average social damages associated with trips that pass through the cordon zones during this period. "Social damages" include both congestion and pollution damages. The second-best optimal cordon prices were calculated using Equation 10 — they reflect both heterogeneity in trip-level externalities, and leakage in time and space.

[12]The NHTS does not report detailed trip start and end locations for states that are not part of the NHTS Add-On program. The trip-level data for New York therefore come from city-specific travel surveys.

[13]Traffic density data for Los Angeles is publicly available through PeMS. Traffic density for NY is courtesy of the NYSDOT Traffic Monitoring Section.

Table 7 — Congestion, Pollution, and Welfare Effects of Peak-Hour Cordon Pricing

	Performance Relative to the First-Best (%)		
Outcome	San Francisco	Los Angeles	New York
Reduction in Total Externalities	29.846	20.908	39.105
Reduction in Congestion	30.741	21.361	39.489
Reduction in Pollution	22.031	16.320	34.938
Welfare Gain	28.370	15.060	42.717

Table 7: This table compares the second-best optimal peak-hour cordon pricing scheme in 3 US cities to an (infeasible) first-best policy where vehicles are charged according to the total marginal damages associated with each trip. "Peak hours" are defined as 6-10am and 3-7pm. Second-best cordon prices are constrained to be uniform during these hours. The four outcomes of interest are total externalities (pollution and congestion), congestion alone, pollution alone, and total welfare (the utility of drivers, in dollars, less total externalities). The figures in this table reflect 600,000 simulated choices using the mixed logit model shows in Table 3.

Table 8 — Congestion, Pollution, and Welfare Effects of Flexible Hourly Cordon Pricing

	Performance Relative to the First-Best (%)		
Outcome	San Francisco	Los Angeles	New York
Reduction in Total Externalities	44.459	36.596	49.358
Reduction in Congestion	45.097	37.044	50.086
Reduction in Pollution	38.975	32.072	41.457
Welfare Gain	51.104	29.392	39.706

Table 8: This table compares the performance of second-best optimal cordon pricing scheme in 3 US cities to an (infeasible) when cordon prices are allowed to vary flexibly between 6am and 7pm. The four outcomes of interest are total externalities (pollution and congestion), congestion alone, pollution alone, and total welfare (the utility of drivers, in dollars, less total externalities). The figures in this table reflect 600,000 simulated choices using the mixed logit model shows in Table 3. The "first-best" is a policy where all trips (regardless of the time of day or whether they pass through the cordon) are charged according to the marginal damages associated with that trip.

Table 9 — Back of the Envelope Welfare Gains From Cordon Pricing

	Welfare Gain Relative to the Status Quo ($ Million)		
Policy	San Francisco	Los Angeles	New York
First-Best	852	1,100	1,477
Second-Best (Peak Only)	246	166	426
Second-Best (Flexible Hourly)	412	323	715

Table 9: This table displays back of the envelope calculations for the total welfare gains under three road pricing policies: 1) The (infeasible) policy where all trips (including those that re-route to avoid a City's cordon) are priced according to marginal congestion and pollution damages; 2) second-best peak hour (6-10 a.m. and 3-7 p.m.) prices (see Table 6); and 3) flexible second-best-optimal prices, which are allowed to vary by hour between 6 a.m. and 6 p.m. The cordon prices in rows (2) and (3) are calculated using Equation 10 — they reflect both heterogeneity in trip-level externalities, and leakage in time and space.

9. Discussion

Cordon prices deviate from first-best policies along two important dimensions: incomplete coverage allows for leakage, and uniform prices cannot reflect the heterogeneity in trip-level damages. I find that these two imperfections are in tension as they apply to optimal road prices. Absent leakage, the correlation between price-responsiveness and trip-level externalities (see Table 3) imply second-best prices that are *above* marginal damages. Absent heterogeneity, leakage to unpriced roads or times of day implies second-best prices that are *below* marginal damages. The results from Figure 14 show that the substitution effect strongly dominates in this setting: optimal prices for San Francisco's cordon zone, for instance, are $2.10 for the morning peak and $2.9 for the evening peak — roughly half of the average social cost for trips that pass through the cordon at those times. Table 6 shows the leakage effect also dominates in New York and Los Angeles. In New York, the second-best optimal cordon prices are about $7 for both the morning and evening peaks, which is below the average social damages associated with cordon trips in each of those periods ($8.18 and $12.64, respectively). In Los Angeles, the optimal a.m. and p.m. peak prices are $3.30 and $4.53, compared to average social damages of $4.87 and $5.72.

Tables 7 and 9 presents the welfare implications of these imperfections. The benefits from optimal peak-hour cordon prices range from $246 million annually in San Francisco to $426 million annually in New York. To put these figures in perspective, the 2021 annual budget of the City of San Francisco is $13.7 billion, and the 2021 annual budget of New York is $88.2 billion. These annual welfare gains are therefore on the order of 0.5 to 2% of city budgets. Tables 7 and 9 also suggest that that cordon pricing is relatively ineffective at reducing congestion and pollution externalities. Optimal peak-hour cordons achieve between 15% (Los Angeles) and 41% (New York) of the welfare gains that would be realized under a first-best policy. Notably, peak-hour pricing policies are less effective at internalizing pollution externalities than they are at internalizing congestion externalities. This reflects the fact that a) congestion externalities represent the majority of the social damages from an average cordon trip and are therefore implicitly more heavily weighted in the optimal tax formula, and b) trip-level pollution externalities are not highly temporally correlated with congestion externalities, as shown in Figures 13 and 12. Together, these findings suggest that peak-hour cordon policies are unlikely to achieve the twin goals of congestion and pollution reduction put forth by many cities.

What (if any) adjustment could improve the performance of the proposed cordon zones in the United States? Relative to a peak-only tolling system, allowing for flexible hourly prices (Table 8) provides sizeable welfare gains:

$146 million in San Francisco, $157 million in Los Angeles, and $286 million in New York). In each city, however, this flexible pricing strategy fails to achieve half of the welfare gains relative to the first-best. Determining the value of spatial expansion of cordon zones or choosing between options for expansion if a more difficult empirical exercise. Sandmo's principle of targeting suggests that if a policymaker chooses to expand a cordon, they should add the roads associated with the largest unpriced externalities, (Sandmo, 1975; Jacobsen, Knittel, Sallee, and Van Benthem, 2020). The theory outlined in Section 2 suggests a modified approach: Prioritize expansion in space based on the b_j terms in Equation 10, which simultaneously take into account direct externalities, heterogeneity, and leakage. In practice, this implies comparing b_j terms across different segments of road. Intuitively, this modified approach to addressing externality taxation may mean that an optimal spatial cordon zone includes outlying roads that are not exceptionally congestion prior to the imposition of a cordon zone, but would become congested if excluded from a city-center cordon.

10. Conclusion

This paper makes three contributions: First, this paper generates the first estimates of optimal cordon prices that account for both pollution and congestion externalities. While optimal prices vary across proposed cordon zones in the US, several themes emerge: Congestion externalities constitute the bulk of marginal damages that determine optimal cordon prices, generally outweighing pollution externalities five- to ten-fold. This finding accords with work by Parry and Small (2005) who suggest that congestion (rather than pollution) is the largest component of an optimal gasoline tax. Additionally, optimal cordon prices tend to be *below* the average social damages associated with trips that cross through a cordon because of externality leakage in time and space. This leakage effect dominates the heterogeneity effect (see Diamond (1973)), which, all else equal, pushes second-best optimal prices above average social damages.

Second, this paper presents the first estimates of the welfare losses that result from imperfections in real-world cordon policies. Back of the envelope calculations suggest that while a second-best peak hour cordon price in San Francisco would produce roughly $200 million dollars worth of welfare gains, this policy would fall short of the first-best policy by $270 million annually. This foregone welfare is significant: $270 million is roughly 2% of the City San Francisco's 2020-2021 Budget ($13.7 billion). The predicted performance of proposed cordons in New York and Los Angeles are qualitatively similar. Notably, among these imperfect policies, the peak-hour cordon zone in New York performs the best (capturing 42% of possible welfare gains), and the peak-hour cordon in Los Angeles performs the worst (capturing just 15% of possible welfare gains). This likely reflects the fact that it is much more difficult to find substitute routes in New York than it is in Los Angeles due to idiosyncratic geography.

Lastly, this paper contributes to public and environmental economics by extending existing models of second best-taxation to simultaneously account for leakage and heterogeneity in externalities. Accounting for these policy imperfections implies subtly different policy prescriptions than the canonical "Principle of Targeting" Sandmo (1975): When externality leakage and externality heterogeneity are present, the policy instrument that generates the largest welfare improvements may not be the tax that best targets the naive average of externalities. Instead, for each good, the optimal instrument balances the magnitude of externality reduction with the damages that would result from leakage. The results in this paper highlight a case where, due to policy imperfections, the optimal policy differs significantly from a tax that best targets the average of consumption externalities. While applying the second-best tax framework outlined in this paper requires detailed information about externalities and consumer demand, the increasing availability of microdata continues to lower the costs for credible estimation of demand systems. This trend, together with the ubiquity of imperfections in externality taxation, suggest that this framework will be useful for future research in settings outside of optimal road pricing.

References

Allcott, Hunt, Benjamin B Lockwood, and Dmitry Taubinsky. 2019. "Regressive sin taxes, with an application to the optimal soda tax." *The Quarterly Journal of Economics* 134 (3):1557–1626.

Anderson, Michael L. 2020. "As the wind blows: The effects of long-term exposure to air pollution on mortality." *Journal of the European Economic Association* 18 (4):1886–1927.

Anderson, Michael L and Maximilian Auffhammer. 2014. "Pounds that kill: The external costs of vehicle weight." *Review of Economic Studies* 81 (2):535–571.

Arnott, Richard, Andre De Palma, and Robin Lindsey. 1990. "Economics of a bottleneck." *Journal of urban economics* 27 (1):111–130.

———. 1993. "A structural model of peak-period congestion: A traffic bottleneck with elastic demand." *The American Economic Review* :161–179.

Auffhammer, Maximilian and Ryan Kellogg. 2011. "Clearing the air? The effects of gasoline content regulation on air quality." *American Economic Review* 101 (6):2687–2722.

Blomquist, Sören, Whitney K Newey, Anil Kumar, and Che-Yuan Liang. 2021. "On bunching and identification of the taxable income elasticity." *Journal of Political Economy* 129 (8):000–000.

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.

Brown, Jennifer, Justine Hastings, Erin T Mansur, and Sofia B Villas-Boas. 2008. "Reformulating competition? Gasoline content regulation and wholesale gasoline prices." *Journal of Environmental economics and management* 55 (1):1–19.

Chetty, Raj, John N Friedman, Tore Olsen, and Luigi Pistaferri. 2011. "Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records." *The quarterly journal of economics* 126 (2):749–804.

City of San Francisco. 2021. "City Performance Scorecards." .

Currie, Janet and Reed Walker. 2011. "Traffic congestion and infant health: Evidence from E-ZPass." *American Economic Journal: Applied Economics* 3 (1):65–90.

Davis, Lucas W. 2008. "The effect of driving restrictions on air quality in Mexico City." *Journal of Political Economy* 116 (1):38–81.

———. 2017. "Saturday driving restrictions fail to improve air quality in Mexico City." *Scientific reports* 7 (1):1–9.

Davis, Lucas W and James M Sallee. 2020. "Should electric vehicle drivers pay a mileage tax?" *Environmental and Energy Policy and the Economy* 1 (1):65–94.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif. 2019. "The mortality and medical costs of air pollution: Evidence from changes in wind direction." *American Economic Review* 109 (12):4178–4219.

Diamond, Peter A. 1973. "Consumption externalities and imperfect corrective pricing." *The Bell Journal of Economics and Management Science* :526–538.

Gibson, Matthew. 2019. "Regulation-induced pollution substitution." *Review of Economics and Statistics* 101 (5):827–840.

Gibson, Matthew and Maria Carnovale. 2015. "The effects of road pricing on driver behavior and air pollution." *Journal of Urban Economics* 89:62–73.

Giuliano, Genevieve. 1992. "An assessment of the political acceptability of congestion pricing." *Transportation* 19 (4):335–358.

Green, Colin P, John S Heywood, and Maria Navarro. 2016. "Traffic accidents and the London congestion charge." *Journal of public economics* 133:11–22.

Green, Colin P, John S Heywood, and Maria Navarro Paniagua. 2020. "Did the London congestion charge reduce pollution?" *Regional Science and Urban Economics* 84:103573.

Green, Jerry and Eytan Sheshinski. 1976. "Direct versus indirect remedies for externalities." *Journal of Political Economy* 84 (4, Part 1):797–808.

Hanna, Rema, Gabriel Kreindler, and Benjamin A Olken. 2017. "Citywide effects of high-occupancy vehicle restrictions: Evidence from "three-in-one" in Jakarta." *Science* 357 (6346):89–93.

Heo, Jinhyok, Peter J Adams, and H Oliver Gao. 2016. "Public health costs of primary PM2. 5 and inorganic PM2. 5 precursor emissions in the United States." *Environmental science & technology* 50 (11):6061–6070.

Hernandez-Cortes, Danae and Kyle C Meng. 2020. "Do environmental markets cause environmental injustice? Evidence from California's carbon market." Tech. rep., National Bureau of Economic Research.

Holland, Stephen P. 2012. "Emissions taxes versus intensity standards: Second-best environmental policies with incomplete regulation." *Journal of Environmental Economics and management* 63 (3):375–387.

Jacobs, Bas and Ruud A De Mooij. 2015. "Pigou meets Mirrlees: On the irrelevance of tax distortions for the second-best Pigouvian tax." *Journal of Environmental Economics and Management* 71:90–108.

Jacobsen, Mark R, Christopher R Knittel, James M Sallee, and Arthur A Van Benthem. 2020. "The use of regression statistics to analyze imperfect pricing policies." *Journal of Political Economy* 128 (5):1826–1876.

Johnson, M Bruce. 1964. "On the economics of road congestion." *Econometrica: Journal of the Econometric Society* :137–150.

Kleven, Henrik J and Mazhar Waseem. 2013. "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan." *The Quarterly Journal of Economics* 128 (2):669–723.

Kleven, Henrik Jacobsen. 2016. "Bunching." *Annual Review of Economics* 8:435–464.

Knittel, Christopher R, Douglas L Miller, and Nicholas J Sanders. 2016. "Caution, drivers! Children present: Traffic, pollution, and infant health." *Review of Economics and Statistics* 98 (2):350–366.

Knittel, Christopher R and Ryan Sandler. 2018. "The welfare impact of second-best uniform-Pigouvian taxation: evidence from transportation." *American Economic Journal: Economic Policy* 10 (4):211–42.

Kopczuk, Wojciech. 2003. "A note on optimal taxation in the presence of externalities." *Economics Letters* 80 (1):81–86.

Kreindler, Gabriel E. 2018. "The welfare effect of road congestion pricing: Experimental evidence and equilibrium implications." *Unpublished paper* .

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

Lehe, Lewis. 2019. "Downtown congestion pricing in practice." *Transportation Research Part C: Emerging Technologies* 100:200–223.

Muller, Nicholas Z and Robert Mendelsohn. 2007. "Measuring the damages of air pollution in the United States." *Journal of Environmental Economics and Management* 54 (1):1–14.

Parry, Ian WH. 2009. "Pricing urban congestion." *Annu. Rev. Resour. Econ.* 1 (1):461–484.

Parry, Ian WH and Kenneth A Small. 2005. "Does Britain or the United States have the right gasoline tax?" *American Economic Review* 95 (4):1276–1289.

Parry, Ian William Holmes. 2002. "Comparing the efficiency of alternative policies for reducing traffic congestion." *Journal of public economics* 85 (3):333–362.

Percoco, Marco. 2016. "The impact of road pricing on accidents: a note on Milan." *Letters in Spatial and Resource Sciences* 9 (3):343–352.

Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *The economic journal* 37 (145):47–61.

Saez, Emmanuel. 2010. "Do taxpayers bunch at kink points?" *American economic Journal: economic policy* 2 (3):180–212.

San Francisco County Traffic Authority. 2021. "Downtown Congestion Pricing Study: Winter 2021 Update." Tech. rep., San Francisco County Traffic Authority.

Sandmo, Agnar. 1975. "Optimal taxation in the presence of externalities." *The Swedish Journal of Economics* :86–98.

———. 1978. "Direct versus indirect Pigovian taxation." *European Economic Review* 7 (4):337–349.

Savidge, Nico. 2021. "Chart: Five ways COVID changed Bay Area traffic." .

Small, Kenneth A, Erik T Verhoef, and Robin Lindsey. 2007. *The economics of urban transportation*. Routledge.

Tonne, Cathryn, Sean Beevers, Ben Armstrong, Frank Kelly, and Paul Wilkinson. 2008. "Air pollution and mortality benefits of the London Congestion Charge: spatial and socioeconomic inequalities." *Occupational and Environmental Medicine* 65 (9):620–627.

Train, Kenneth E. 2009. *Discrete choice methods with simulation*. Cambridge university press.

Verhoef, Erik, Peter Nijkamp, and Piet Rietveld. 1995. "Second-best regulation of road transport externalities." *Journal of transport economics and policy* :147–167.

Vickrey, William S. 1963. "Pricing in urban and suburban transport." *The American Economic Review* 53 (2):452–465.

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.

Zhang, Wei, C-Y Cynthia Lin Lawell, and Victoria I Umanskaya. 2017. "The effects of license plate-based driving restrictions on air quality: Theory and empirical evidence." *Journal of Environmental Economics and Management* 82:181–220.

Zhong, Nan, Jing Cao, and Yuzhu Wang. 2017. "Traffic congestion, ambient air pollution, and health: Evidence from driving restrictions in Beijing." *Journal of the Association of Environmental and Resource Economists* 4 (3):821–856.

Appendix

A. Theory Appendix

A.1. Substitution with Many Goods

Setup: A representative consumer chooses quantities of M goods, $(h_1, ..., h_M)$ and a numeraire, z. Each non-numeraire good has an associated externality, ϕ_m. A policymaker can choose tax levels for goods $j \in [1, J]$ where $J < M$. I assume goods $m \notin [1, J]$ are un- or under-taxed.

The consumer's problem: An agent maximizes utility over M goods $(h_1, ..., h_M)$ and a numeraire good z.

$$max\{U(h_1, ..., h_M) + z\} \quad s.t. \tag{26}$$

$$(p_1 + \tau_1)h_1 + (p_J + \tau_J)h_J + p_{J+1}h_{J+1} + ... + p_M h_M + z \leq I \tag{27}$$

The first-order conditions for an interior solution to the consumer's problem are:

$$U_j = \lambda(p_j + \tau_j) \quad \forall j \in [1, J] \tag{28}$$

$$U_m = \lambda(p_m) \quad \forall m \notin [1, J] \tag{29}$$

$$\lambda = 1 \tag{30}$$

The planner's problem: I assume that the planner seeks to maximize aggregate welfare, which is the utility of the representative consumer less the aggregate social cost of consumption, $\sum_1^M \phi_m h_m$. The planner's choice variables are tax levels $\tau_1 ... \tau_J$, which are applied to the taxable goods $j \in [1, J]$.

$$\begin{aligned} max\{U(h_1, ..., h_M) + z - \sum_1^M \phi_m h_m\} \; st. \\ p_1 h_1 + ... + p_N h_N + z \leq I \end{aligned} \tag{31}$$

Assuming an internal solution, first-order condition wrt p_j (where $j \in [1, J]$) is:

$$0 = \frac{\partial h_j}{\partial p_j}[U_j - \phi_j - p_j] + \sum_{k \neq j}^{M} \frac{\partial h_k}{\partial p_j}[U_k - \phi_k - p_k] \tag{32}$$

Plugging in the consumer's first order conditions and solving for τ_m...

$$0 = \frac{\partial h_j}{\partial p_j}[\tau_j - \phi_j] + \sum_{k \neq j}^{J} \frac{\partial h_k}{\partial p_j}[\tau_k - \phi_k] + \sum_{l=J+1}^{M} \frac{\partial h_l}{\partial p_j}[\phi_l] \tag{33}$$

$$\tau_j = \phi_j + \frac{1}{\frac{\partial h_j}{\partial p_j}}(\sum_{k \neq j}^{J} \frac{\partial h_k}{\partial p_j}[\phi_k - \tau_k] + \sum_{l=J+1}^{M} \frac{\partial h_l}{\partial p_j}\phi_l) \tag{34}$$

This intermediate results is intuitive. Holding fixed all taxes other than τ_j, the optimal value for this final tax is its externality, ϕ_m, minus a term that captures the extent to which consumers switch to other goods, and the level of unpriced externality of those goods.

Identifying the optimal tax level for *all* taxable goods requires solving J equations simultaneously:

$$\tau_j + \frac{1}{\frac{\partial h_j}{\partial p_j}}(\sum_{k \neq j}^{J} \frac{\partial h_k}{\partial p_j}\tau_k) = \phi_j + \frac{1}{\frac{\partial h_j}{\partial p_j}} \sum_{l=1}^{M} \frac{\partial h_l}{\partial p_j}\phi_l \tag{35}$$

This gives us J equations, each linear in the J tax levels:

$$\alpha_1^j \tau_1 + ... + \alpha_k^j \tau_k + ... + \alpha_J^j \tau_J = b_j \qquad \forall j \in [1, J] \tag{36}$$

Where α_k^j and b_m are defined as:

$$\alpha_k^j = \frac{\frac{\partial h_k}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \tag{37}$$

$$\beta_j = \phi_j + \sum_{l=1}^{M} \frac{\frac{\partial h_l}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}}\phi_l \tag{38}$$

The α and β terms have an intuitive interpretation. α_k^j is the share of the reduction in overall consumption of good j that shifts to good m as a results of an increase in the price of good j. β_j is the overall reduction in externalities that results form the increase in the price of good j; this consists of a direct component, ϕ_j plus a (negative) leakage term, $\sum_{l=1}^{M} \frac{\partial h_l}{\partial p_j} / \frac{\partial h_j}{\partial p_j}\phi_l$.

This system can be written as:

$$\begin{bmatrix} \alpha_1^1 & ... & \alpha_J^1 \\ & ... & \\ \alpha_1^j & ... & \alpha_J^J \end{bmatrix} \begin{bmatrix} \tau_1 \\ ... \\ \tau_J \end{bmatrix} = \begin{bmatrix} b_1 \\ ... \\ b_J \end{bmatrix} \tag{39}$$

$$\boldsymbol{A\tau} = \boldsymbol{b} \tag{40}$$

$$\boldsymbol{\tau} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{41}$$

A.2. Heterogeneity and Substitution

Setup: N Heterogeneous consumers choose between M externality-generating goods and a numeraire, z. I denote individual i's consumption of good m as h_i^m. Each individual has an exogenous income μ_i. I assume that each consumer's utility is a function of their consumption of these M goods and a quasilinear numeraire, as well as other's consumption of these goods (which generate externalities and decrease i's utility): $U_i(h_1^1, ...h_1^M, ..., h_i^1, ...h_i^M, ...h_N^1, ...h_N^M) + z_i$.

As in section 2.3, a policymaker can choose tax levels for goods $j \in [1, J]$ where $J < M$. I assume goods $m \notin [1, J]$ are un- or under-taxed. I denote τ^j as the tax on good j.

The consumer's problem: Agent i maximizes utility over M goods $(h_i^1, ..., h_i^M)$ and their consumption of the numeraire good z_i.

$$\begin{aligned} max\{U_i(h_1^1, ...h_1^M, ..., h_i^1, ...h_i^M, ...h_N^1, ...h_N^M) + z_i\} \; st. \\ (p^1 + \tau^1)h_i^1 + (p^J + \tau^J)h^J + p^{J+1}h_i^{J+1} + ... + p^M h_i^M + z_i \leq \mu_i \end{aligned} \tag{42}$$

The first-order conditions for this problem are:

$$\begin{aligned} \frac{\partial U_i}{\partial h_i^j} &= \lambda(p^j + \tau^j) \qquad \forall j \in [1, J] \\ \frac{\partial U_i}{\partial h_i^m} &= \lambda(p^m) \qquad \forall m \notin [1, J] \\ \lambda &= 1 \end{aligned} \tag{43}$$

The planner's problem: I assume that the planner seeks to maximize aggregate welfare, $\sum_1^N (U_i + z_i)$. The planner's choice variables are tax levels $\tau^1 ... \tau^J$, which are applied to the taxable goods $j \in [1, J]$.

$$\begin{aligned} & max\{\sum_i^N (U_i(h_1^1, ...h_1^M, ..., h_i^1, ...h_i^M, ...h_N^1, ...h_N^M) + z_i) \\ st. \qquad & (p^1)\sum_i^N h_i^1 + ... + (p^J)\sum_i^N h_i^J + (p^{J+1})\sum_i^N h_i^{J+1} + ... + (p^M)\sum_i^N h_i^M + \sum_i^N z_i \leq \sum_i^N \mu_i \end{aligned} \tag{44}$$

Assuming an internal solution, first-order condition wrt p^j (where $j \in [1, J]$) is:

$$0 = \sum_{i=1}^N \frac{\partial U_i}{\partial h_i^l}\frac{\partial h_i^l}{\partial p_j} + \sum_{i=1}^N \sum_{g \neq i}^N \frac{\partial U^i}{\partial h_g^1}\frac{\partial h_1^g}{\partial p_j} + ... + \frac{\partial U^i}{\partial h_g^M}\frac{\partial h_g^M}{\partial p_j} - p^1 \sum_i \frac{\partial h_i^1}{\partial p_j} - ... - p^M \sum_i \frac{\partial h_1^M}{\partial p_j} \tag{45}$$

Plugging in the consumer's first order conditions and solving for τ_j...

$$\tau_j = \frac{\sum_{i=1}^N \sum_g^N (\frac{\partial U^i}{\partial h_g^1}\frac{\partial h_g^1}{\partial p_j} + ... + \frac{\partial U^i}{\partial h_g^M}\frac{\partial h_g^M}{\partial p_j})}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} + \frac{\sum_{k \neq j}^J \frac{\partial h_i^k}{\partial p_j}\tau_k}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} \tag{46}$$

This expression for the optimal level of a given tax is equivalent to the equation for substitutes with homogeneous damages where each of the marginal damages have been replaced by a "Diamond" term which accounts for heterogeneity.

B. Calculating Emissions Externalities

This section details the process of estimating emissions externalities for each trip in the FasTrak dataset.

The California Emissions Factor (EMFAC) fleet database reports average vehicle emissions rates (measured in grams per mile) by county. These data are stratified by vehicle fuel type, vehicle vintage, and vehicle travel speed. The EMFAC database reports the following pollutant species: particulate matter (PM2.5, or PM), nitrogen oxides (NO_x), nitrous oxide (N_2O), reactive organic compounds (ROC), ammonia (NH_3), carbon dioxide (CO_2), sulfur oxides (SO_2), and methane (CH_4). The data underlying EMFAC aggregates reflect state vehicle registrations and data from the California Bureau of Automotive Repair's (BAR) Smog Check database. For each FasTrak trip, I assign emission factors for each pollutant based on the average travel speed for that trip (see Appendix D) and the county where the FasTrak device is registered. The total emissions of any pollutant is the estimated emissions *rate* for that trip multiplied by the trip *length*.

To convert trip-level emissions to costs, I use social cost estimates from two sources. For local pollutants, I use damages predicted by the EAISUR model (Heo, Adams, and Gao, 2016), which combines a state-of-the-art chemical transport model together with estimates from the economics and epidemiology literatures to predict the cost of emitting pollution in different areas of the United States. For global pollutants, I use social damages from the US EPA. These pollutant values are listed in Table 13.

Table 13 — Social Costs of Vehicle Pollution in San Francisco

Pollutant	**Damage ($/Ton)**
$PM_{2.5}$	772,000
SO_2	65,800
NO_x	24,200
NH_3	1,24,000
CO_2	51
CH_4	1,500
N_2O	18,000
ROC	2,392

Table 10: This table display the social costs of emitting 1 ton of various pollutants in San Francisco. Estimates of local pollutants (PM2.5, or PM), nitrogen oxides (NO_x), nitrous oxide (N_2O), reactive organic compounds (ROC), ammonia (NH_3), sulfur oxides (SO_2)) reflect annual averages from the EAISUR model (Heo, Adams, and Gao, 2016). Global pollutants (carbon dioxide (CO_2) and methane (CH_4)) are values used by the US EPA.

Table 13 — Social Costs of Vehicle Pollution in Los Angeles

Pollutant	**Damage ($/Ton)**
$PM_{2.5}$	1,270,000
SO_2	44,750
NO_x	52,750
NH_3	825,750
CO_2	51
CH_4	1,500
N_2O	18,000
ROC	2,392

Table 11: This table display the social costs of emitting 1 ton of various pollutants in Los Angeles. Estimates of local pollutants (PM2.5, or PM), nitrogen oxides (NO_x), nitrous oxide (N_2O), reactive organic compounds (ROC), ammonia (NH_3), sulfur oxides (SO_2)) reflect annual averages from the EAISUR model (Heo, Adams, and Gao, 2016). Global pollutants (carbon dioxide (CO_2) and methane (CH_4)) are values used by the US EPA.

Table 13 — Social Costs of Vehicle Pollution in Los Angeles

Pollutant	**Damage ($/Ton)**
$PM_{2.5}$	1,270,000
SO_2	44,750
NO_x	52,750
NH_3	825,750
CO_2	51
CH_4	1,500
N_2O	18,000
ROC	2,392

Table 12: This table display the social costs of emitting 1 ton of various pollutants in Los Angeles. Estimates of local pollutants (PM2.5, or PM), nitrogen oxides (NO_x), nitrous oxide (N_2O), reactive organic compounds (ROC), ammonia (NH_3), sulfur oxides (SO_2)) reflect annual averages from the EAISUR model (Heo, Adams, and Gao, 2016). Global pollutants (carbon dioxide (CO_2) and methane (CH_4)) are values used by the US EPA.

Table 13 — Social Costs of Vehicle Pollution in New York City

Pollutant	**Damage ($/Ton)**
$PM_{2.5}$	$1,270,000$
SO_2	$44,750$
NO_x	$52,750$
NH_3	$825,750$
CO_2	51
CH_4	$1,500$
N_2O	$18,000$
ROC	$2,392$

Table 13: This table display the social costs of emitting 1 ton of various pollutants in New York City. Estimates of local pollutants (PM2.5, or PM), nitrogen oxides (NO_x), nitrous oxide (N_2O), reactive organic compounds (ROC), ammonia (NH_3), sulfur oxides (SO_2)) reflect annual averages from the EAISUR model (Heo, Adams, and Gao, 2016). Global pollutants (carbon dioxide (CO_2) and methane (CH_4)) are values used by the US EPA.

C. Bunching Estimator

Additional Bunching Results

Table 14 — Bunching Estimator for Scheduling Costs (Shifting Later)

Parameter	Estimate
Fraction Unresponsive (a)	0.87357
Excess Mass at Notch (B)	0.00136
Baseline Density at Notch, h0	0.00059
Mean Schedule Cost without Friction ($/hour)	87.51086
Mean Schedule Cost accounting for Frictions ($/hour)	11.06393
Mean Schedule Cost accounting for Frictions and Travel Time ($/hour)	15.49765

Table 14: Rows 1-3 of this table show estimates of parameters used to infer scheduling costs from the additional density of trips just after the end of peak-hour pricing on San Francisco's Bay Bridge (equation 22). Rows 4 and 5 show estimates of scheduling costs themselves. In Row 4, I calculate the naive average scheduling cost for bunchers under the assumption that there are no optimization frictions. In row 5, I use the estimated fraction of non-responsive individuals from row 1 to account for optimization frictions.

Table 15 — Bunching Estimator for Scheduling Costs (Shifting Earlier)

Parameter	Estimate
Fraction Unresponsive (a)	0.76058
Excess Mass at Notch (B)	0.00208
Baseline Density at Notch	0.00019
Mean Schedule Cost without Friction ($/hour)	18.65659
Mean Schedule Cost accounting for Frictions ($/hour)	4.46673
Mean Schedule Cost accounting for Frictions and Travel Time ($/hour)	6.25671

Table 15: Rows 1-3 of this table show estimates of parameters used to infer scheduling costs from the additional density of trips just after the end of peak-hour pricing on San Francisco's Bay Bridge (equation 22). Rows 4 and 5 show estimates of scheduling costs themselves. In Row 4, I calculate the naive average scheduling cost for bunchers under the assumption that there are no optimization frictions. In row 5, I use the estimated fraction of non-responsive individuals from row 1 to account for optimization frictions.

Figure 19 — Travel Times in the Vicinity Price Notches

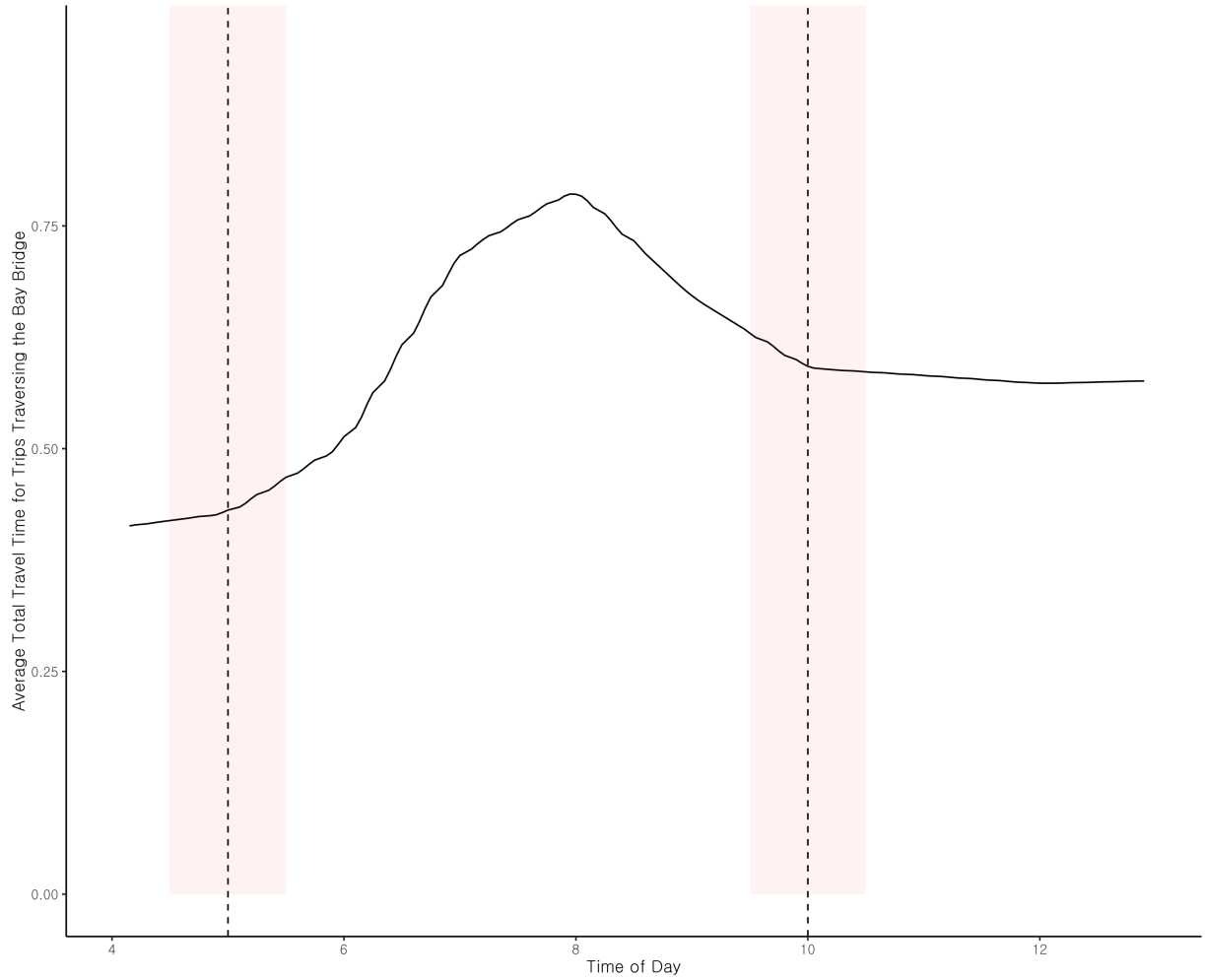


Figure 19: This figure plots average travel times for trips traversing the Bay Bridge during the morning hours. Average travel times are calculating by identifying all drivers that primarily use the Bay Bridge and using TomTom Historic Traffic Stats to calculate travel times for each individual, for each hour of day. The red shaded area represent the rough range where individuals adjust in response to the imposition of peak-hour pricing. The relatively flat profile of travel times in the price notch neighborhood suggests that the first-order decision facing drivers who travel at this time of day is between price and scheduling costs.

D. Estimating Travel Times

Travel times, T_i, are not directly observed for FasTrak trips, and therefore must be inferred. In this appendix, I describe the process for inferring $T(h_i, r_i)$ for each trip in each individual's choice set.

The choice set of any individual consists of all bridges, {*Dumbarton Bridge, San Mateo Bridge, Bay Bridge, Richmond Bridge*} at all times of day, {4.0, 4.2, ..., 22}. A *trip* in this choice set constitutes a bridge-time pair. I estimate travel times for each trip in each individual's choice set in three steps:

Step 1: Infer the distribution of endpoints. The FasTrak tolling data include information about the bridges used, as well as the home zip code associated with each FasTrak device. Before calculating travel times using historic traffic data, I must make inferences about the missing endpoints for each driver. To do so, I use survey data from the 2010-2012 California Household Travel Survey (CHTS). These two surveys constitute a representative sample of Bay Area commuters, an contain detailed information on the driving habits of respondents. To generate a probability distribution of "work" endpoints for each individual, I subset the CHTS survey data to trips that match based on home city and bridge used. The Bay Area is relatively unique in that it is a large metropolitan area that consists of many small cities. The 29 "cities" that serve as termini for travel time estimation are plotted in Figure 20.

Step 2: Calculate travel times. I use TomTom's Historic Traffic Stats to calculate the travel times. For each device, I calculate a travel time between the device's home city and each of the endpoints assigned positive probability in Step 1. Importantly, I estimate travel times for both trips that were taken, as well as counterfactual trips that used a different bridge or were taken at a different hour of day.

Step 3: Aggregate travel times by bridge and time of day. Lastly, I collapse the distribution of possible travel times within each hour by the probability weights from Step 1. The result is a data set that contains estimated travel times for each trip taken by each device, as well as the travel times that a driver would have faced for each trip had they taken it at a different hour of day or using a different bridge.

FIGURE 20 — TOMTOM TRAFFIC SEGMENTS

Figure 20: This figure plots the coverage of the TomTom historic traffic stats data (in red) together with the 29 most populous cities in the Bay Area. These roads were selected using Google Maps suggested driving points between the origin and destination cities.

E. Congestion Pricing and Accidents

In this appendix, I outline the rationale for excluding accident externalities from this analysis.

In a manner similar to congestion and pollution externalities, the decision to drive imposes external accident risk on other drivers. Anderson and Auffhammer (2014) show that this externality relies crucially on vehicle weight, and exceeds congestion and pollution externalities for the average US driver.

Large accident externalities for the average US driver, however, may not translate to higher optimal cordon prices because of differences in the risk of rural vs. urban driving. Empirical studies of the impact of congestion charges on accidents suggest that the value of accident reductions are several orders or magnitude smaller than pollution and congestion externalities. Green, Heywood, and Navarro (2016), for example, find that the London cordon zone reduced overall accidents by 35%, and fatal accidents by 25 to 35%. Because of the relatively low number of fatal auto-related deaths in London, however, the authors value these safety improvements at just £28 million annually. For comparison, Leape (2006) estimates the congestion benefits from London's cordon zone were estimated at £230 million annually. Similarly, Percoco (2016) finds that while Milan's Cordon Zone reduced overall traffic accidents by 16 to 18%, there was no detectable impact on fatal accidents. Valuations of associated benefits are therefore dominated by the roughly $3 billion in reduced pollution and congestion externalities (Gibson and Carnovale, 2015).

The relatively small impact of congestion pricing on severe accidents may reflect the fact that many of the main risk factors severe traffic accidents — high traffic speeds, drinking and driving, and nighttime driving — are not well targeted by cordon zones. Relatedly, driving in cities in the US and Europe tends to be relatively safe overall, making it straightforward to put bounds on the accident-related benefits that may accrue from congestion pricing.

In San Francisco, for example, there are 20 to 30 fatal accidents (including pedestrian fatalities) each year (City of San Francisco, 2021). Under a $10 million value of a statistical life, reducing traffic fatalities in San Francisco by 30% would be worth roughly $90 million dollars —– an order of magnitude smaller than my estimated of the combined congestion and pollution benefits associated with cordon pricing in San Francisco. All indicators suggest that a cordon zone would fall well short of this mark. During 2020, for example, the number of traffic fatalities (31) did not fall amid the 30% pandemic-related decrease in Bay-Area traffic (City of San Francisco, 2021; Savidge, 2021).

Together, these pieces of evidence suggests that it is unlikely that accounting for accident externalities would substantively change the conclusions in this paper.

F. Interactions with Existing Taxes and Revenue Requirements

In this appendix, I cover the interaction between road pricing and existing environmental policies, as well as the literature on whether governmental revenue requirements impact the optimal Pigouvian tax.

F.1. Accounting for Existing Environmental Taxes

Broadly speaking, in the presence of existing Pigouvian taxes the optimal level for an *additional* tax covers the difference between the marginal damages associated with consumption and the existing corrective tax. It is therefore important to account for existing environmental policies that act as a tax on driving when calculating optimal Pigouvian road prices.

There are a number of State and Federal policies that regulate vehicle-related local pollution emissions in California. These policies largely fall into two categories: Tailpipe emissions regulations (e.g., catalytic converter

requirements) and fuel content regulations (e.g., volatile organic compound regulations). Below, I use a simple model to demonstrate that these two types of policies have different implications for designing an additional tax to internalize remaining externalities associated with driving. Regulations that impact vehicle costs should *not* be taken into account when calculating optimal road prices. The costs of fuel content regulations, however, should be subtracted from road prices to the extent that these regulations lead to higher per-mile driving prices.

Existing policies that impact vehicle cost:

Consider a representative household with exogenous income I that consumes two goods, driving x and a quasi-linear numeraire good z. Driving is associated with an externality, $\phi(a)$. The per-mile magnitude of this externality can be abated (a) on the assembly line at cost $c(a)$. I assume that ϕ_a and c_a are each differentiable, with $c'(a) > 0$ and $\phi`(a) < 0$. The planner's problem is to choose an abatement level, a and a diving level x to maximize total welfare:

$$W = u(x) + z - \phi(a) \cdot x - c(a) \quad \text{s.t.} \quad I \geq z - p \cdot x$$

The Lagrangian associated with this maximization problem is:

$$\mathcal{L} = u(x) + z - \phi(a) \cdot x - c(a) + \lambda(I - z - p \cdot x)$$

The first-order conditions for an interior solution to this problem are:

$$\lambda = 1$$
$$u'(x) = \phi(a) + p$$
$$\phi'(a)x = c'(a)$$

These conditions imply that the planner equates marginal abatement costs and marginal abatement benefits, and (separately) equates marginal driving costs and marginal driving benefits. The fact that abatement costs do not enter directly into the first order condition for x implies that if a is set at some exogenous level, the policymaker would ignore the abatement cost when choosing the optimal level of driving, only weighing the utility of driving against the externalities that remain after abatement. I therefore ignore the costs of environmental policies that impact vehicle prices (e.g., requirements for catalytic converters) when calculating the level of "unpriced" externalities for drivers.

Existing policies that impact fuel cost:

Now consider the same consumer model, but the per-mile magnitude of this externality can be abated by altering fuel content at cost $c(a) \cdot x$. That is, the total abatement cost now depends on the amount of driving, x.

Again consider a policymaker who maximizes total social welfare, W:

$$W = u(x) + z - (\phi(a) - c(a)) \cdot x; \quad \text{s.t.} \quad I \geq z - p \cdot x$$

The Lagrangian associated with this maximization problem is:

$$\mathcal{L} = u(x) + z - (\phi(a) - c(a)) \cdot x + \lambda(I \geq z - p \cdot x)$$

The first-order conditions with respect to x and a are:

$$\lambda = 1$$
$$u'(x) = \phi(a) + c(a) + p$$
$$\phi'(a) = c'(a)$$

As above, these first-order conditions imply that the planner equates marginal abatement costs and marginal abatement benefits, and equates marginal driving costs and marginal driving benefits. The crucial difference in this case is that the marginal cost of driving now includes abatement costs. As a result, the social planner will still weight these costs when setting optimal road prices. I therefore include the costs of environmental policies that impact fuel prices (e.g., fuel content regulation) when calculating the level of "unpriced" externalities for drivers.

When calculating optimal cordon prices in this paper I use a cost of 12 cents per gallon for gasoline for the pre-existing cost of fuel content regulations. This is the midpoint of estimates of the costs of California's fuel content regulations from Auffhammer and Kellogg (2011) (who in turn rely on estimates from Brown, Hastings, Mansur, and Villas-Boas (2008) and the California Air Resources Board), converted to 2021 dollars.

F.2. Accounting for Government Revenue Requirements

The stylized models above raise the question of whether *any* policy that increases the per-mile cost of driving about the competitive equilibrium should be accounted for when calculating optimal road prices. Work by Kopczuk (2003) and Jacobs and De Mooij (2015) suggests that optimal taxation and Pigouvian taxation are separable problems: The calculation of optimal road prices should not take into account taxes that exist as a result of governments balancing the distortions of various revenue sources.

As noted by Jacobs and De Mooij (2015), however, this argument relies on the fact that the marginal cost of public funds is one in an optimal tax system. If the marginal cost of public funds is *not* one, then the optimal second-best Pigouvian tax could be higher or lower than a tax set equal to marginal social damages. Absent strong evidence that the marginal cost of public funds is above or below one, I assume that the marginal cost of public funds is one in the San Francisco Bay Area, and therefore do not adjust optimal road prices to reflect their interactions with the tax system. As anecdotal evidence of this assumption, note that California state and local ballot initiatives frequently feature direct votes on taxation, bond issuance, and spending decisions. It is plausible that this low barrier to public finance reform allows California's tax code to reflect citizen's preferences for public goods and redistribution more accurately than do tax codes regions without ballot initiatives.

G. External Validity

The appropriateness of using of the travel demand model estimated using data from the San Francisco Bay Area (see Section 7) to cordon pricing in other cities depends on whether trips taken in other cities are similarly substitutable, and whether similar correlations between trip-level externalities and price responsiveness are present. In this appendix, is use data from the National Household Transportation Survey to investigate these relationships for two other US cities — New York and Los Angeles — that are currently considering implementing congestion pricing.

Broadly, NHTS data suggest that the relevant relationships in each of these cities are similar to those in San Francisco. Drivers appear similarly able to shift trips temporally. Figure 21, for example, shows that similar fractions of drivers report flexible work schedules in each of these cities. Figure 22 shows that likelihood of a given trip being flexible varies in New York and Los Angeles in a manner similar to the within-day variation in San Francisco. Figures 23 through 25 provide suggestive evidence that the way that externalities generated by driving — congestion and pollution — vary with price responsiveness in New York and Los Angeles is similar to the way that these externalities vary with price responsiveness in San Francisco. In each city, drivers who "agreed" or "strongly agreed" that gasoline prices impacted their decision to drive were modestly more likely to drive an older, more polluting vehicle. Similarly, drivers that report being more responsive to gas prices report driving along more congested routes, measured as the difference in reported commute time with vs. without traffic.

Figure 21 — Schedule Flexibility by Metro Area

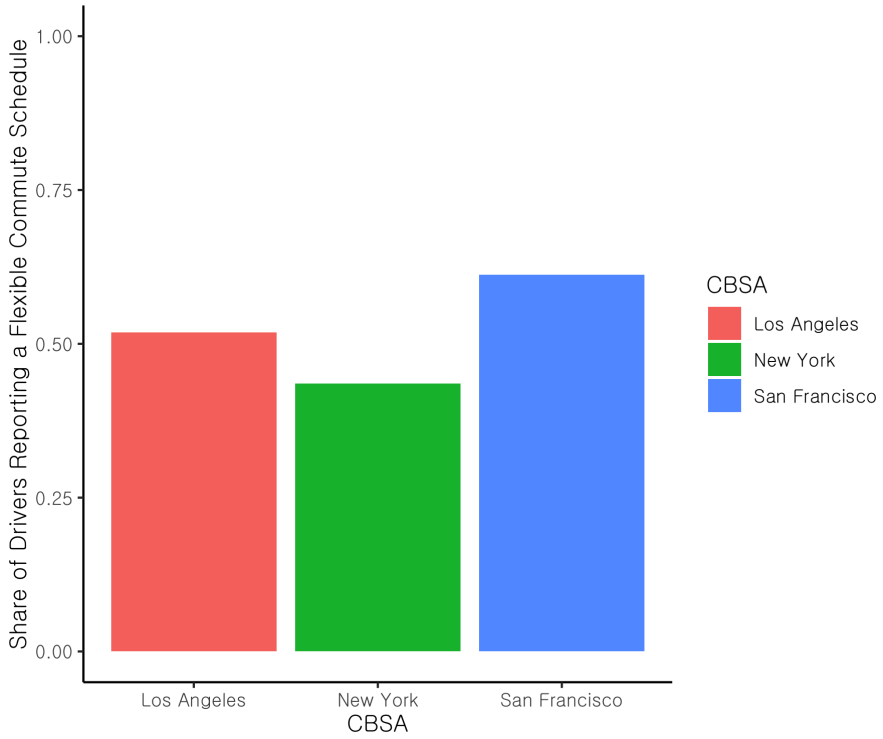


Figure 21: This figure plots the share of drivers who report having a flexible work schedule by metro area, according to the 2017 National Household Transportation Survey.

Figure 22 — Schedule Flexibility by Time of Day

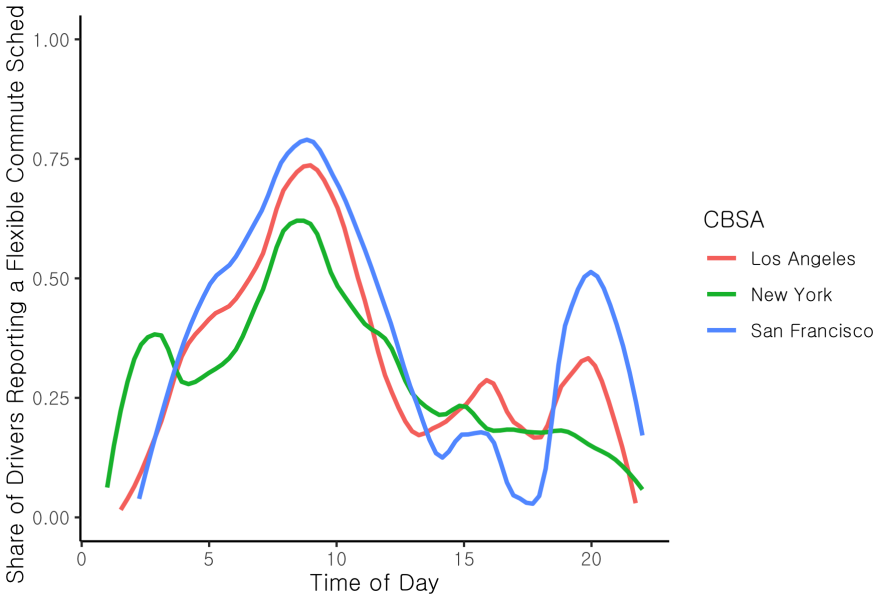


Figure 22: This figure plots the share of drivers who report having a flexible work schedule by time of day and metro area, according to the 2017 National Household Transportation Survey.

Figure 23 — Emissions Factors vs. Gas Price Responsiveness

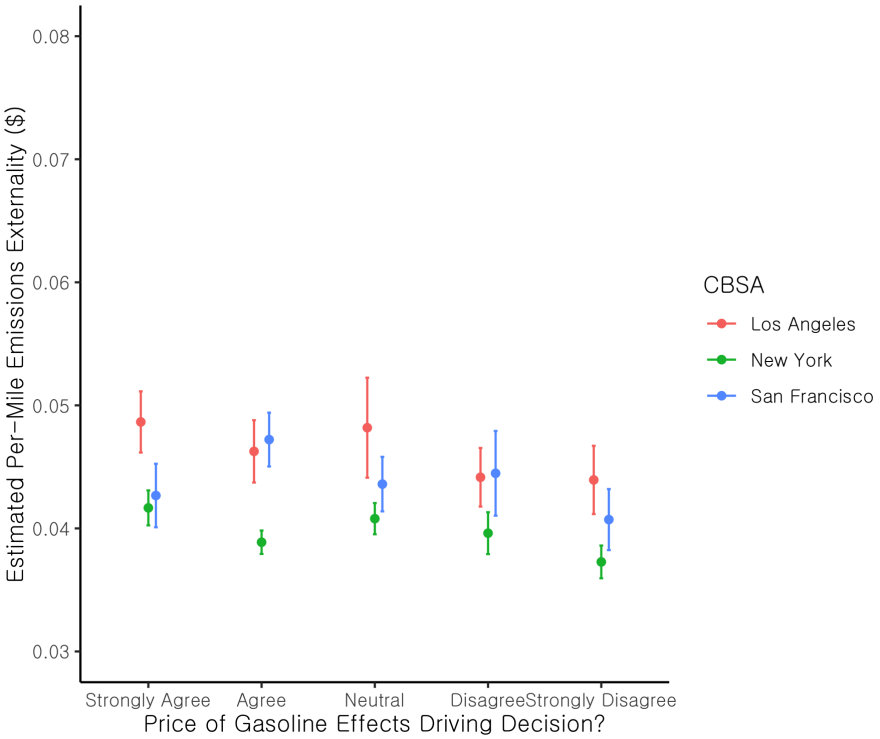


Figure 23: This figure plots estimates emissions factors of vehicles in the 2017 National Household Transportation Survey against vehicle owners' self-reported responsiveness of travel demand with respect to gasoline prices. Emissions factors reflect vehicle age and fuel type.

Figure 24 — Vehicle Age vs. Gas Price Responsiveness

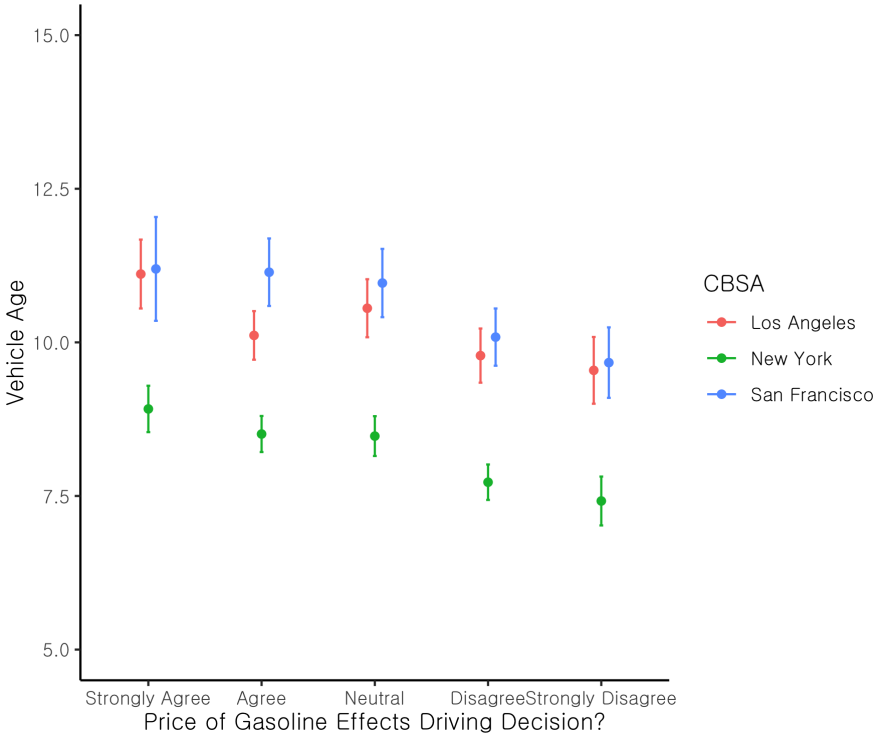


Figure 24: This figure plots vehicle age against vehicle owners' self-reported responsiveness of travel demand with respect to gasoline prices using data from the 2017 National Household Transportation Survey.

Figure 25 — Congestion vs. Gas Price Responsiveness

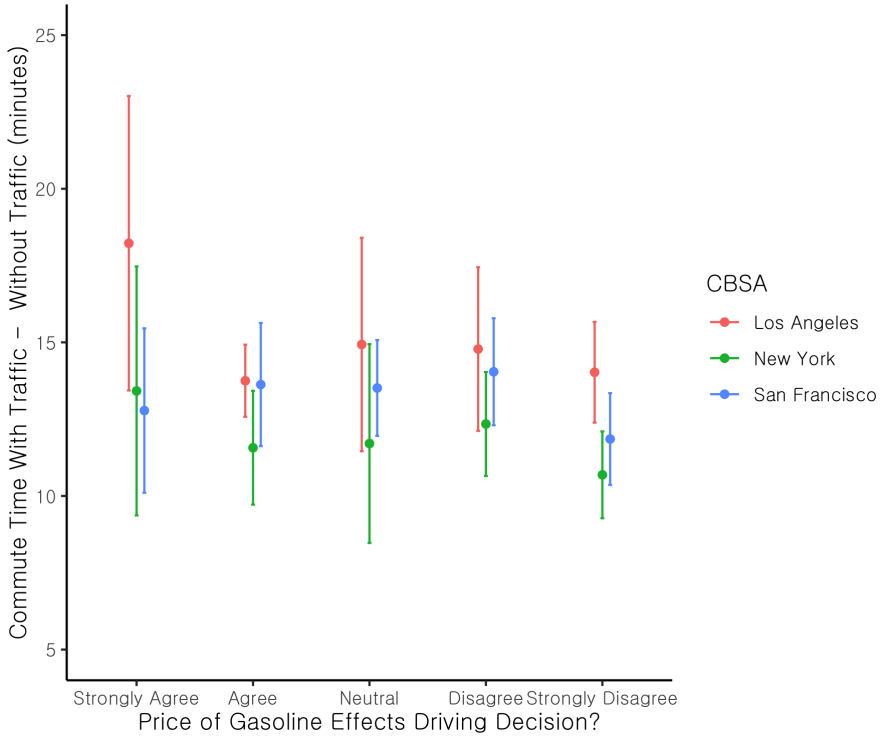


Figure 25: This figure plots self-reported gasoline price responsiveness against the amount of time a driver reports loosing to traffic during their commute for drivers in the 2017 National Household Transportation Survey.