

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

Matthew Tarduno[†]

November 5, 2021

JOB MARKET PAPER

[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 drivers according to the social damages associated with each trip. In practice, road pricing often takes the form of *cordon zones* — regions in the center of a city where road users are charged for entry. These pricing schemes deviate from the first-best policy in two key ways: 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 information about (i) 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 administrative 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). Due to the blunt nature of cordon pricing, these policies are relatively ineffective at internalizing congestion and pollution externalities. In these three cities, I estimate that second-best optimal cordon prices recover 15 to 40% of the welfare gains that would be achieved under an ideal 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 pricing.

[†]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, Severin Borenstein, James Sears, Jenya Kahn-Lang, Marshall Blundell, Matthew Suandi, Elif Tasar, Eva Lyubich, Sara Johns, and seminar participants at UC Berkeley, Williams College, NC State's Camp Resources, and OSWEET 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	3
2	Theory: Externality Taxation Under Heterogeneity and Leakage	5
3	A Discrete Choice Model of Driver Behavior	10
4	Natural Experiment: Traffic Tolling in the San Francisco Bay Area	11
5	Data	14
6	Empirical Strategy	17
7	Results	22
8	Second-Best Optimal Cordon Prices	26
9	Discussion	36
10	Conclusion	37

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 little insight on how to implement road pricing in practice, especially given that real-world policy instruments differ significantly from the first-best policies prescribed by economists.

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 often rejected on privacy grounds ([Lehe, 2019](#); [Giuliano, 1992](#)). Consequently, city-wide road pricing 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 congestion and pollution externalities across trips that all enter the cordon. 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. I then generate empirical estimates of how drivers would respond to road pricing, and use these estimates together with formulas derived from the theoretical framework to calculate second-best cordon prices.

The second-best pricing framework I build stipulates a set of parameters necessary for calculating second-best road prices accounting for both leakage and imperfect pricing (i.e., many vehicle trips with different externalities are charged the same price). Calculating optimal prices requires information about (i) 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 use 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 administrative microdata from the region's electronic tolling system to estimate a discrete choice model of driving 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](#)). The structure of this discrete choice model simplifies the information required to apply the second-best tax framework (items (i)-(iii), above). Namely, it allows me to populate a substitution matrix between alternative driving times and routes based on a small number of parameters that describe driving choices.

I use this model of driving 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, resulting in second-best optimal prices that are below the average social damages associated with trips that enter the cordon. 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, peak-hour cordon pricing performs poorly relative to the (infeasible) Pigouvian prescription. The second-best optimal road pricing scheme in San Francisco, for example, 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, and are therefore poorly targeted by peak-hour congestion prices.

To conclude, I investigate the prospects for improving cordon pricing policies. Allowing a policymaker to set a fixed schedule of hourly prices between 6 a.m. and 7 p.m. generates sizable welfare gains relative to a cordon policy constrained to charge prices only during peak hours. I estimate that these welfare gains range from \$146 million annually in San Francisco to \$286 million annually in New York. In each city, however, a cordon zone with second-best hourly prices would realize less than half of the welfare gains possible under first-best prices.

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 \$2.20 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, as noted by 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). Calculating optimal cordon prices through a second-best tax framework explicitly accounts for these considerations.

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 combines two canonical models of second-best pricing: the “Diamond” model (Diamond, 1973), 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, the 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 energy markets, for example, the externalities associated with electricity generation differ based on the location of powerplants (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 key parameters in the transportation economics literature (Vickrey (1963), Arnott, De Palma, and Lindsey (1990), Arnott, De Palma, and Lindsey (1993)) and an important driver of the theoretical welfare gains from congestion pricing (Kreindler, 2018). 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 to understand scheduling in many other 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 discrete choice model of driving 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 driving demand, and Section 5 covers the data. In Section 6, I describe the empirical strategy that I use to estimate the model of driving 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 Leakage

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 leakage (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 leakage), and then present a model that can be applied to instances where heterogeneity and leakage simultaneously prevent the implementation of the first-best.

2.1. Heterogeneity

For practical or legal reasons, policymakers are often constrained to apply a uniform corrective price to a good where the consumption externalities associated with that good are not uniform. In cordon zones, for example, drivers typically face a single charge for daytime trips, or a toll that charges one price for peak-hour trips, and a lower price for off-peak trips.

Under these pricing schemes, many trips that generate different externalities will be charged the same price.

¹ Sources of congestion heterogeneity include the total 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, and trip length.

Diamond (1973) characterizes the second-best optimal uniform tax on a good which generates heterogeneous externalities when consumed by different agents: The optimal tax 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:

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

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

$$\tau^* = \frac{-\sum_h \sum_{h \neq i} \frac{\partial U^h}{\partial \alpha_i} \alpha'_i}{\sum_h \alpha'_h} \quad (1)$$

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.

2.2. Leakage

Legal, political, or practical constraints often prevent policymakers from pricing all related externality-producing goods. Cordon prices, for example, price only trips that pass over the cordon's boundary, leaving trips that avoid the cordon unpriced.

Green and Sheshinski (1976) show that 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 the taxable good at its marginal damages, less a term that is increasing in the substitutability of the two goods, and increasing in the marginal damages of consuming the untaxed 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, z :

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

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

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

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

The second-best optimal price balances the direct social damages associated with consumption of the taxable good (ϕ_x), with the leakage-associated social damages that result from an increase in the price of the taxable good ($\frac{dy/dp_x}{dx/dp_x} \phi_y$). In this paper, I will refer to the first term in this expression ($\frac{dy/dp_x}{dx/dp_x}$) as the *leakage share* between x and y .

As noted by [Davis and Sallee \(2020\)](#), equation 2 reflects separability insights from [Kopczuk \(2003\)](#): The second-best tax on x is the sum of a) direct damages associated with the consumption of this good, an b) a term that captures interactions between x and existing market distortions.

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. Leakage with Many Goods

Before characterizing second-best optimal taxes under both heterogeneity and leakage, I first extend the two-good model in Section 2.2 to the case of many (homogeneous) externality-generating goods, some of which are untaxed. This problem is a generalization of the two and three-good direct vs. indirect taxation problems presented in [Green and Sheshinski \(1976\)](#) and [Sandmo \(1978\)](#), and provides intuition useful for understanding the model with heterogeneity presented in Section 2.4.

Setup: A representative consumer chooses quantities of M goods, (h_1, \dots, h_M) and a numeraire, z . Each non-numeraire good has an associated (homogeneous) externality, ϕ_m that is linear in the consumption of m . A policymaker can choose tax levels τ_j for goods $j \in \{1, \dots, J\}$ where $J < M$. I assume goods $k \notin \{1, \dots, J\}$ 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:

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

This intermediate results is a generalization of the two-good case. 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:

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

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

$$a_1^j \tau_1 + \dots + a_l^j \tau_l + \dots + a_J^j \tau_J = b_j \quad \forall j \in [1, J] \quad (4)$$

Where a_l^j and b_j are defined as:

$$a_l^j = \frac{\frac{\partial h_l}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \quad (5)$$

$$b_j = \phi_j + \sum_{m=1}^M \frac{\frac{\partial h_m}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \phi_l \quad (6)$$

The a and b terms have intuitive interpretations. a_l^j is the share of the reduction in overall consumption of good j that shifts to good l as a result of an increase in the price of good j . That is, each a term is a leakage share between taxable goods. b_j is the overall reduction in externalities that results from the increase in the price of good j ; this consists of a direct component, ϕ_j , plus the sum of leakage terms: $\sum_{m=1}^M \frac{\partial h_m}{\partial p_j} / \frac{\partial h_j}{\partial p_j} \phi_l$, which are negative if j is a normal good and m is a substitute for j .

This system can be written compactly as:

$$\begin{bmatrix} a_1^1 & \dots & a_J^1 \\ \vdots & \ddots & \vdots \\ a_1^J & \dots & a_J^J \end{bmatrix} \begin{bmatrix} \tau_1^* \\ \vdots \\ \tau_J^* \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_J \end{bmatrix}$$

$$\mathbf{A}\boldsymbol{\tau} = \mathbf{b} \quad (7)$$

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

$$\boldsymbol{\tau} = \mathbf{A}^{-1} \mathbf{b} \quad (8)$$

Equation 8 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 Leakage

Finally, I characterize second-best taxes where a) there are many externality-generating products, b) policy-makers 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 (Herrnstadt, Parry, and Siikamäki, 2015).

Lastly, as I introduce heterogeneity, it is worth noting that I assume that the social planner acts to maximize aggregate welfare, as in Diamond (1973). The formulae that follow do not account for redistributive preferences

²Note that this substitution matrix contains cross-price consumption *derivatives* and not cross-price consumption *elasticities*. \mathbf{A} contains 1’s along the diagonal; when all j goods are substitutes, the off-diagonal terms of \mathbf{A} fall in the closed interval $[0, -1]$.

— heterogeneity is included in the model to reflect the implications of differences in externalities, rather than understand how externality taxation interacts with inequality aversion.³

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): $U_i(h_1^1, \dots, h_1^M, \dots, h_i^1, \dots, h_i^M, \dots, h_N^1, \dots, h_N^M) + z_i$.

As in Section 2.3, a policymaker can choose tax levels for goods $j \in \{1, \dots, J\}$ where $J < M$. I assume goods $m \notin \{1, \dots, J\}$ 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:

$$\tau_j = \frac{\sum_{i=1}^N \sum_g^N (\frac{\partial U^i}{\partial h_g^1} \frac{\partial h_g^1}{\partial p_j} + \dots + \frac{\partial U^i}{\partial h_g^M} \frac{\partial h_g^M}{\partial p_j})}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} + \frac{\sum_{k \neq j}^J \frac{\partial h_i^k}{\partial p_j} \tau_k}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} \quad (9)$$

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:

$$\begin{bmatrix} a_1^1 & \dots & a_J^1 \\ a_1^2 & \dots & a_J^2 \\ \vdots & & \vdots \\ a_1^J & \dots & a_J^J \end{bmatrix} \begin{bmatrix} \tau_1^* \\ \vdots \\ \tau_J^* \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_J \end{bmatrix} \quad (10)$$

Where a_l^j and b_j are defined as:

$$a_l^j = \frac{\sum_{i=1}^N \frac{\partial h_i^l}{\partial p^j}}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} \quad (11)$$

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

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.

These are considerable information requirements. In what follows, I demonstrate how to use the structure of discrete choice modeling to reduce the dimensionality of this problem. Specifically, rather than estimating how each driver substitutes between each possible trip, I use a discrete choice model over routes and times of day to populate the substitution matrix of options facing drivers based on the attributes of those trips.

³See Allcott, Lockwood, and Taubinsky (2019) for a characterization optimal corrective taxation with incomplete instruments and social preferences for redistribution.

3. A Discrete Choice Model of Driver Behavior

The theory outlined in Section 2 stipulates 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 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 their ideal arrival time incurs disutilities γ_e and γ_l per minute, respectively. Drivers also incur disutility α from each minute spent commuting. Utility is thus:

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

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

$$\{h_i^*, r_i^*\} = \arg \max_{h_i, r_i} \{u(h_i, r_i)\} \quad (14)$$

To clarify the mapping between this discrete choice model and the optimal tax formula (Equation 10), a “good” (h^j in the notation used in Section 2) is a trip taken on a given *route* at a given *time of day*: $\{h, r\}$. Typical cordon zones have discrete spatial and temporal cutoffs.⁴ The possibility of leakage reflects the ability of drivers to adjust trips in time (h) and space (r) to avoid tolls. Heterogeneity in externalities results from the fact that trips that enter a cordon zone during the same time of day are charged the same price, but 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). To estimate the relationship between idiosyncratic externalities and price-responsiveness, I allow β (the coefficient on price) to vary across externality quantiles during estimation.

The value of estimating this discrete choice model is that it greatly reduces the number of parameter estimates required for applying the optimal tax formula outlined in Section 2. For any choice set (e.g., cordon vs. non-cordon routes at various times of day), equation 13 implies a matrix of own and cross-price elasticities between choices, which reflect model primitives ($\alpha_e, \gamma_e, \gamma_l, \beta$) and trip attributes ($T, p, \text{time late}, \text{and time early}$). In Section 4 through 7, I use tolling microdata to recover estimates of each of these parameters using a mixed logit model. I also apply a bunching estimator to the introduction of peak-hour pricing in the Bay Area to produce separate estimates of scheduling parameters, γ_e and γ_l .

⁴The 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 San Francisco would only charge drivers for trips during peak hours (6-10 am and 3-7 pm).

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

I use administrative tolling data from the San Francisco Bay Area together with revisions to regional bridge tolls to estimate the model of driving demand outlined in Section 3.

4.1. Bay Area Bridge Tolls

FasTrak is an electronic tolling 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.⁵ 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 map shows the four San Francisco Bay Area bridges used to estimate driver responses to toll prices in this paper. The *Richmond Bridge* connects Richmond and the eastern Bay Area to San Rafael and Marin County. The *Bay Bridge* connects Oakland to San Francisco. The *San Mateo Bridge* connects Hayward to San Mateo. The *Dumbarton Bridge* connects Fremont to Palo Alto. Each of these bridges charges drivers for westbound trips (as detailed in Figure 2).

4.2. Variation in Toll Prices

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.⁶ 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.

⁵Drivers 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.

⁶Between 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 or more axles.

In this paper, I leverage several changes in the tolling structure that occurred on July 1, 2010 to identify the parameters necessary to calculate optimal road prices. In 2009, the Bay Area Toll Authority (BATA) adopted Resolution 90, which increased the base prices for passenger vehicles from \$4 to \$5 beginning on July 1, 2010, and established peak-hour pricing on the Bay Bridge (detailed below). This intertemporal variation in toll prices is plotted in Figure 2.

FIGURE 2 — VARIATION IN PASSENGER VEHICLE BRIDGE TOLLS

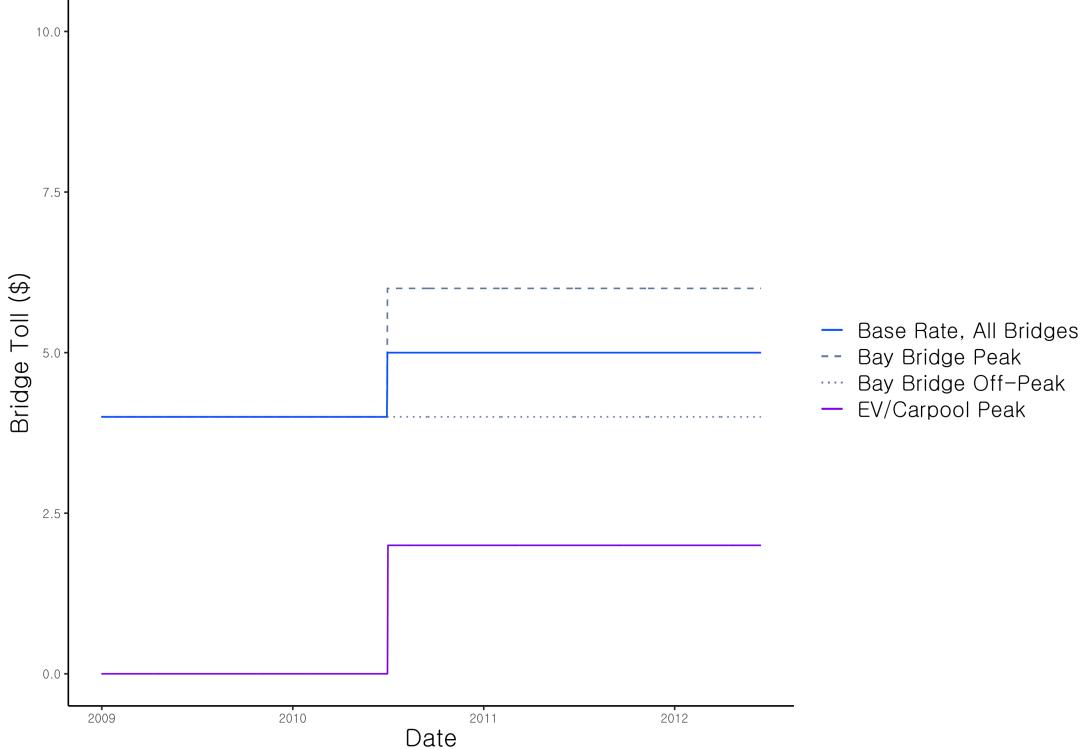


Figure 2: This figure shows Bay Area bridge tolls between 2009 and 2012 for passenger vehicles. Prices are uniform across bridges, with the exception of the Bay Bridge, which connects San Francisco and Oakland. Beginning in 2010, passenger vehicles crossing the Bay Bridge faced a two-dollar difference between peak and off-peak prices. The peak (\$6) and off-peak (\$4) prices are plotted above as dotted and dashed lines, respectively. EV and carpool trips were free on all bridges prior to 2010. Beginning in July of 2010, EV/carpool trips were charged the base rate (\$5 on the San Mateo, Dumbarton, and Richmond Bridges; \$4 on the Bay Bridge), 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 (which connects San Francisco and Oakland) beginning on July 1, 2010. Passenger vehicles crossing westbound through the Bay 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 (henceforth *off-peak*) remained at the pre-2010 price of \$4.

Prior to July 1, 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 1, 2010 and January 1, 2019.

FIGURE 3 — PEAK PRICING ON BAY AREA BRIDGES

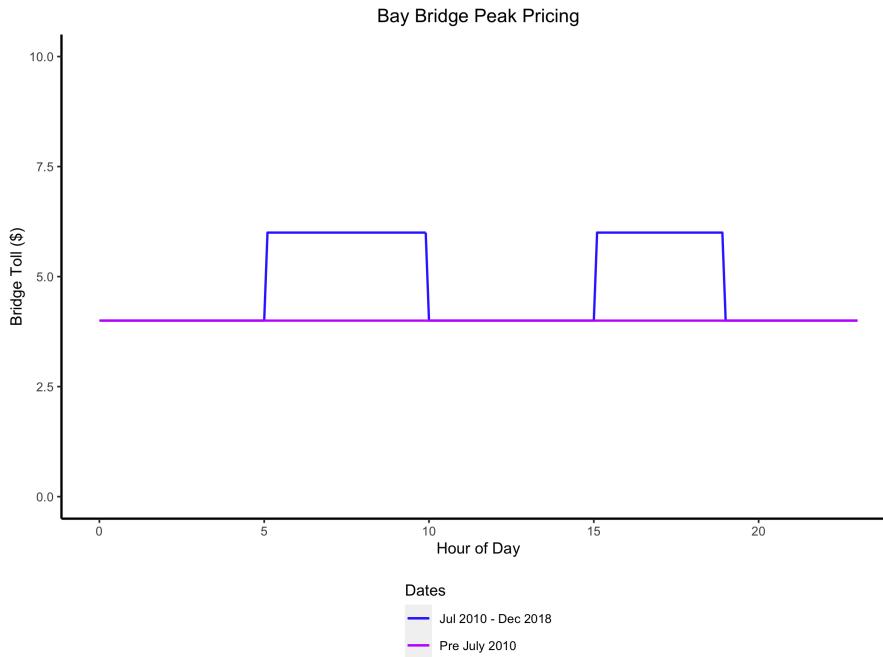


Figure 3: This figure displays peak-hour pricing schemes for passenger vehicles (vehicles with two axles) on California's Bay Bridge, which connects San Francisco and Oakland. Beginning on July 1, 2010, passenger vehicles crossing westbound on weekdays during peak hours (between 5 a.m. and 10 a.m., or 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 charging the base rate for passenger vehicles (\$4 for pre-2010 and \$5 for July 2010 - December 2018).

Foreman (2016) uses reduced-form approaches to provide valuable estimates of the responses of Bay Area drivers to this change in bridge prices. The number of vehicle trips during peak hours on the Bay Bridge decreased by 6 to 8% (400 to 550 vehicles per hour) following the imposition of peak hour pricing. Travel during off-peak hours on the Bay Bridge increased by 4 to 20% (225 to 400 vehicles per hour). Point estimates suggest the \$1 increase on the San Mateo and Dumbarton bridges led to modest decreases in bridge use (15 to 48 vehicles per hour). Notably, crossings on the San Mateo Bridge *increased* by 100 to 200 vehicles (around a 5%) during peak hours, implying that some drivers switched from the Bay Bridge its closest substitute in response to the peak-hour price difference across routes.

To summarize this variation in road prices in this empirical setting, the 2010 revision to bridge tolls in the San Francisco Bay Area replaced uniform prices with prices that varied across bridges and times of day. Reduced-form analyses of this policy suggest that drivers responded to these pricing by reducing the overall number of trips, as well as shifting their trips in time and space. In the following sections, I use this variation in prices together with microdata on driver choices to estimate the model of personal vehicle travel described in Section 3.

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 administrative microdata from the FasTrak tolling system with historic travel time data purchased from TomTom’s *Historic Traffic Stats* database.

FasTrak Toll Data: I use administrative microdata from the FasTrak tolling system to create a panel of individual-level driving choices. These microdata record any transactions that occurred on the four trans-bay bridges between January 1, 2009 and July 1, 2019. A single observation in this data set includes the date, time, and location of the vehicle crossing, as well as the vehicle class (axle number), 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. Roughly 40% of observations that use a FasTrak device also list the home 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 include only devices with a valid (Bay Area) zip code. 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 FasTrak devices and 1,078,044 bridge crossings.

These sample restrictions reflect the information requirements of the discrete choice model of driving demand. Recall that this model specifies driver utility as a function of trip attributes: travel time, time late or early, and price. Zip code information is necessary for assigning travel times to vehicle trips based on the distance between households and bridges. The restrictions based on the frequency of trips reflects the need to infer ideal arrival times for drivers. For FasTrak devices associated with daily commuters, ideal arrival times can be inferred based on bridge-crossing times prior to July 1, 2010 (detailed below). Drivers that infrequently use bridges, or that use bridges many times a day, are not well-described by the discrete choice model I employ in this paper, as it is unclear how to assign these trips an ideal arrival time and trip termini. While imposing these sample restrictions comes at the cost comprehensiveness, estimating the discrete choice model provides a distinct benefit relative to a reduced-form approach: For any given choice set (e.g., driving options subject to cordon prices) the structure of the discrete choice model directly implies the substitution parameters required for calculating optimal prices.

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 Traffic Stats 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 12 months before and the 12 months 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 reconstruction

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 E.

Ideal Arrival Times: Ideal arrival times, h_i^A in equation 13, are not directly observed, and therefore must be inferred from 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 1, 2010 and July 1, 2010, plus the weighted average travel time between the bridge toll plaza and each of the possible endpoints for that driver.

For an illustrative example, consider a driver who exclusively uses the *Bay Bridge* during the pre-period, and who most commonly crosses this bridge at 9 in the morning. A trip taken by this individual that crosses the bridge at 9 a.m. would be assigned a value of zero for *time late* and *time early*. A trip taken by this individual that crosses the bridge at 10 a.m. would be assigned a value of *time late* of 1, plus any difference in expected after-bridge travel time between 9 a.m. and 10 a.m.

Lastly, it is worth noting that pre-period bridge crossing times may not indicate actual ideal crossing times if within-day traffic conditions provide sufficient incentive for drivers to shift their trips in time to reduce overall commute times. The estimates of scheduling elasticities that I recover from responses to peak-hour pricing on the Bay Bridge, however, are inconsistent with this type of strategic scheduling. If Bay Area drivers have schedule costs low enough to induce them to strategically reschedule trips in the absence of peak-hour pricing, a much higher portion of drivers should have responded to the imposition of peak-hour pricing by rescheduling trips.

5.2. Externalities

Although data on trip-level externalities is not necessary for estimating a model of driving 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. Although most estimates of per-mile externalities in the economics literature suggest that accident externalities constitute a significant portion of the overall social costs of driving (Parry and Small, 2005; Anderson and Auffhammer, 2014), empirical evidence suggests that the social benefits from reduced accidents in cordon zones are an order of magnitude smaller than the benefits associated with reduced congestion and air pollution (Green, Heywood, and Paniagua, 2020). Broadly, this empirical evidence reflects the fact that the type of driving curtailed by cordon pricing — slow, daytime trips in city centers — results in relatively few fatal traffic accidents. I provide further discussion of the relationship between accidents and optimal cordon prices in Appendix H.

Congestion Externalities: Congestion externalities vary significantly in space and time. 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. These traffic sensors are mapped in Figure 13. 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)⁷ 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). 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 estimates of the relationship between idiosyncratic externalities and price responsiveness.

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

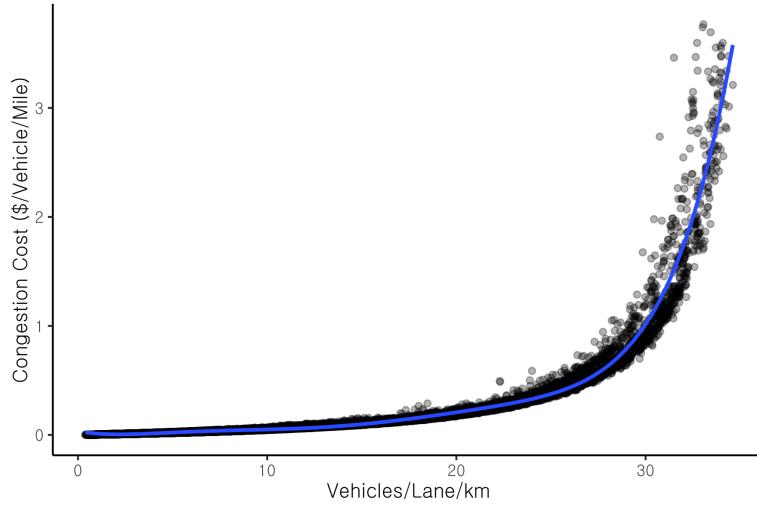


Figure 4: Congestion costs reproduced from [Yang, Purevjav, and Li \(2020\)](#), who exploit variation in traffic density generated by Beijing's driving restriction to estimate the relationship between traffic density and speed. The original results are presented in Yuan/Vehicle/km. I convert these values to dollars by a) converting currencies, and b) replacing the Beijing-specific value of time from (50% of the average wage rate in Beijing) with a \$20 value of travel time, which reflects San Francisco-specific estimates from [Goldszmidt et al. \(2020\)](#).

Emissions Externalities: Fuel combustion and brake wear in passenger vehicles generates several air pollutants. These include “global” pollutants like CO₂ 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), 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 vintag⁸, and vehicle speed.⁹

⁷The 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 the Beijing-specific value of time from (50% of the average wage rate in Beijing) with a \$20 value of travel time, which reflects research by [Goldszmidt, List, Metcalfe, Muir, Smith, and Wang \(2020\)](#).

⁸Older 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 vehicle's lifetime

⁹Vehicle speed impacts emissions through engine efficiency and the intensity of brake wear.

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 to the EASIUR model of local pollution damages. See Appendix C for details on individual pollutant costs.

FIGURE 5 — POLLUTION EXTERNALITIES AT VARIOUS SPEEDS

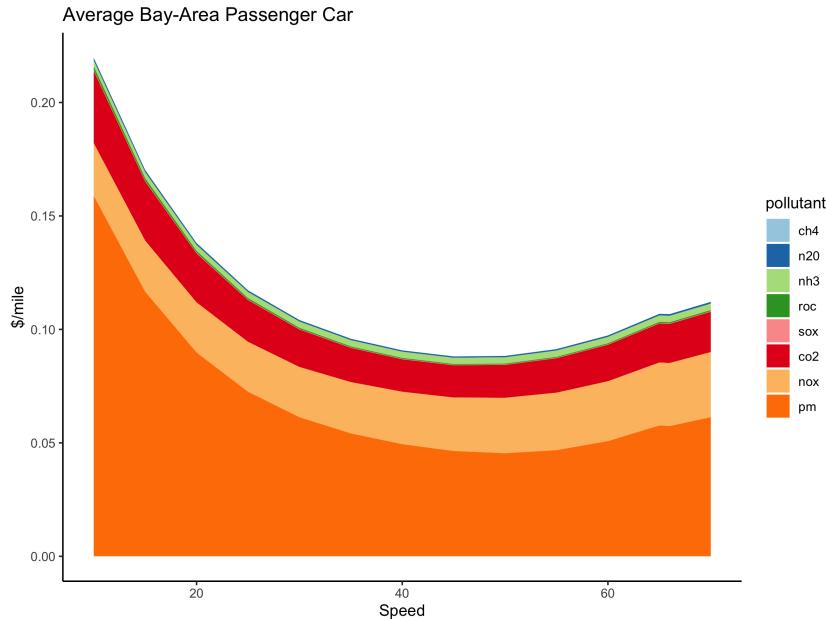


Figure 5: This figure plots per-mile pollution externalities at various speeds for an average passenger vehicle in the Bay Area. These costs reflect VMT-weighted average emissions factors (in grams/mile) of different pollutants at different speeds reported by California's Emissions Factor Model (EMFAC). The EMFAC emissions factor estimates reflect state DMV and smog check data. To convert these emissions factors to per-mile costs, I multiply the emission factor for each pollutant by the corresponding 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 CO₂ and \$1,500 per ton of CH₄, 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.

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 mixed logit regression. 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 Regressions

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 discrete choice model of driving demand outlined in Section 3.

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

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 in 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, which is detailed in Section 4. Peak-hour pricing on the Bay Bridge constitutes a potential threat to identification through reverse causality. 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 error term $\epsilon_{h,r,i}$. If, however, this high demand was the result of factors unobserved by the researcher that make peak-hour travel on the Bay Bridge attractive, 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 address this threat to identification, I also estimate an instrumental variables regression where *post* acts as an instrument for *price*, which leverages only the level shift in prices to estimate the coefficient on price.

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. The variation in total travel times in response to the 2010 change in toll prices is negligible Foreman (2016). 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:

$$\varepsilon_{\{h^j, r^k\}, \{h^l, r^m\}} = \begin{cases} \beta p(h^l, r^m)(1 - s_{\{h^l, r^m\}}), & \text{if } i = l \wedge j = m \\ \beta p(h^l, r^m)s_{\{h^l, r^m\}}, & \text{otherwise} \end{cases} \quad (16)$$

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, ordinary 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 (“mixed”) logit regression. This regression estimates a joint distribution of coefficients (θ) which implies idiosyncratic pairwise substitution parameters between trip options:

$$\varepsilon_{\{h^j, r^k\}, \{h^l, r^m\}} = \frac{p(h^l, r^m)}{s_{\{h^j, r^k\}}} \int \beta s_{\{h^j, r^k\}}(\theta) s_{\{h^l, r^m\}}(\theta) f(\theta) d\theta \quad (17)$$

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 for the estimation of 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 costs and perfect control over when they cross a bridge that charges different tolls during peak and off-peak hours. A “buncher” is a driver who would cross the bridge during peak hours in the absence of peak-hour pricing, but who would adjust their travel time to just avoid the extra toll in a world with peak-hour pricing. For the *marginal* buncher, the utility from the lower price is equal to the scheduling costs of adjusting their trip to cross outside of peak hours. Equation 18 shows this indifference condition in terms of structural parameters. For simplicity, I examine the case of a driver who faces a decision of whether or not to shift their trip earlier:

$$\underbrace{\beta \Delta p}_{\text{Benefit from shifting}} = \underbrace{\gamma_e \Delta h}_{\text{Cost of shifting}} \quad (18)$$

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 time that the marginal buncher would have to adjust their trip in order to cross the bridge before peak hours:

$$\gamma_e = \frac{\beta \Delta p}{\Delta h} \quad (19)$$

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

$$\gamma_e = \frac{\beta \Delta p + \alpha \Delta T}{\Delta h} \quad (20)$$

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 19 and 20 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 more willing to shift their trips further from their ideal travel time for a given level of compensation. A wider density gap therefore implies lower scheduling costs.

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 frictions may prevent drivers from perfectly optimizing (Kleven, 2016). In this setting, these ‘frictions’ may reflect inattentiveness (Finkelstein (2009), for example, finds that toll prices are less salient for drivers with automatic toll tags) or the inability to perfectly time bridge crossings due to traffic shocks.

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:

$$B = \int_{\gamma_e}^{\gamma_e + \Delta h} (1-a)f_0(h)dh \simeq (1-a)f_0(h^*)E[\Delta h] \quad (21)$$

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 21 for Δh and plugging into Equation 20 yields the bunching estimator:

$$\gamma_e = \frac{\beta \Delta p}{B / ((1-a)f_0(h^*))} \quad (22)$$

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

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

In all bunching estimates, I use a \$20 value of travel time, which reflects San Francisco specific findings from Goldszmidt et al. (2020). I also present estimates of scheduling parameters that ignore time savings (equation 22) in Appendix D.

FIGURE 6 — THE RELATIONSHIP BETWEEN SCHEDULING COSTS AND BUNCHING

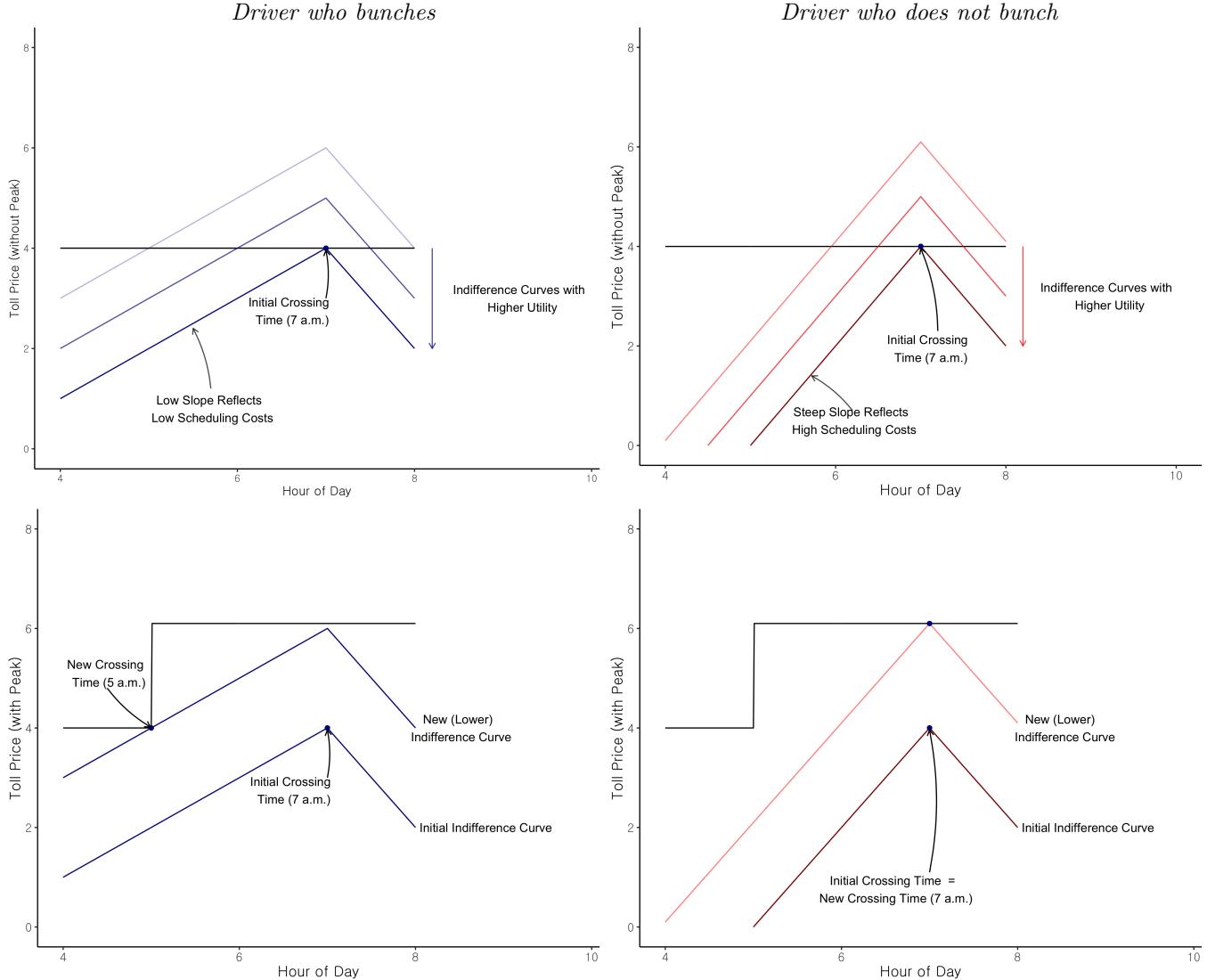


Figure 6: This figure illustrates the relationship between scheduling costs and bunching behavior in peak-hour toll schemes, as predicted by the discrete choice model outlined in Section 3. For expositional ease, this figure plots the case where travel times are constant throughout the day. The triangular shape of the indifference curves 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 right two panes, I plot indifference curves (red) of a driver with *high* scheduling costs, who does not shift their trip in response to peak pricing. In the left two panes, I plot indifference curves of a diver with *low* scheduling costs, who does shift their trip in response to peak pricing. 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 discrete choice model of driving demand (equation 13) estimated using the empirical approaches outlined in Section 6.

7.1. Logit Regression Results

Table 1 presents the results from equation 15, estimated via multinomial logistic regression. The point estimates from this regression imply that on average, drivers are indifferent between saving roughly \$11 and saving an hour of travel time; they are indifferent between saving roughly \$6 arriving an hour early, and they are indifferent between saving roughly \$4 arriving an hour late.

Table 10 and Figure 14 show results of estimating equation 15 via mixed logit. Allowing for heterogeneity in the logit parameters produces results that are qualitatively similar to the results in Table 1. In Table 2, 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. The results in Table 2 suggest that price elasticity and idiosyncratic externalities are positively correlated: drivers that travel longer distances at more congested times and places are more price responsive on average than are drivers who take shorter, less-congested trips.

Table 1 — A DISCRETE CHOICE MODEL OF DRIVING DEMAND

Variable	Specification			
	(1)	(2)	(3)	(4)
Travel Time (\$/hr)	12.626 (1.926)	12.575 (1.912)	11.266 (1.76)	9.659 (1.384)
Time Early (\$/hr)	6.062 (0.307)	6.043 (0.305)	5.514 (0.282)	2.171 (0.366)
Time Late (\$/hr)	4.447 (0.206)	4.433 (0.205)	4.028 (0.19)	2.163 (0.326)
Price	1.000 (1.002)	1.000 (1.001)	1.000 (0.91)	1.000 (0.175)
Day of Week FE		Yes	Yes	Yes
Bridge FE			Yes	Yes

Table 1: Results from Equation 15, a discrete choice model where drivers choose over routes and times of day, estimated using 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 . *Time late* is analogously defined. *Price* is the toll that driver i would incur by traveling via route r at hour h . As the coefficient on price is normalized to 1, the other coefficients can be interpreted as costs in dollars per hour. Columns (1) through (3) show results of a standard logit regression. Column (4) presents results of an instrumental variables regression where an indicator for post July 1, 2010 acts as an instrument for *price*. Standard errors in all regressions are clustered at the individual and zip code levels. All values are in 2010 dollars.

Table 2 — MIXED LOGIT WITH PRICE ELASTICITIES BY EXTERNALITY QUARTILE

variable	mean	sd
time early	-1.247 (0.113)	0.488
time late	-3.409 (0.235)	2.078
travel time	-5.371 (0.53)	1.57
price (first externality quartile)	-0.016 (0.556)	
price (second externality quartile)	-0.148 (0.105)	
price (third externality quartile)	-0.285 (0.108)	
price (fourth externality quartile)	-0.126 0.115	

Table 2: Results from Equation 15, a discrete choice model where drivers choose over routes and times of day, estimated using a random coefficients (“mixed”) logit model. 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 . *Time late* is analogously defined. *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. Two-way standard errors are clustered at the individual and zip code levels.

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 range from \$6 to \$15 per hour.

Figure 7 plots the difference between the density of trips by time of day before vs. after the imposition of peak-hour pricing for the 5 a.m. price notch on San Francisco's Bay Bridge. The bunch in the density of trips prior to 5 a.m. (which does not exist in the data prior to July 1, 2010) is consistent with a model of driving demand where drivers are willing to shift their trips in response to price incentives, but scheduling costs a) prevent all drivers from doing so, and b) lead drivers that do shift to adjust their travel time by the minimum amount necessary to receive the incentive.

Figures 15 and 16 plot the frequency of vehicle trips of before versus after the imposition of peak hour pricing for all hours of day. Qualitatively, the bunches appear to be most pronounced during the early morning (5 a.m.) and early afternoon (3 p.m.) price notches. Intuitively, this suggests that it is less costly to arrive early than arrive late for both morning and evening trips.

Using equation 23, I estimate that during morning commute hours, the marginal driver 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 driver is roughly indifferent between saving \$9 being an hour early, and indifferent between saving \$13 and being an hour late. These estimates are summarized in Table 3

FIGURE 7 — BUNCHING IN RESPONSE TO PEAK-HOUR PRICING

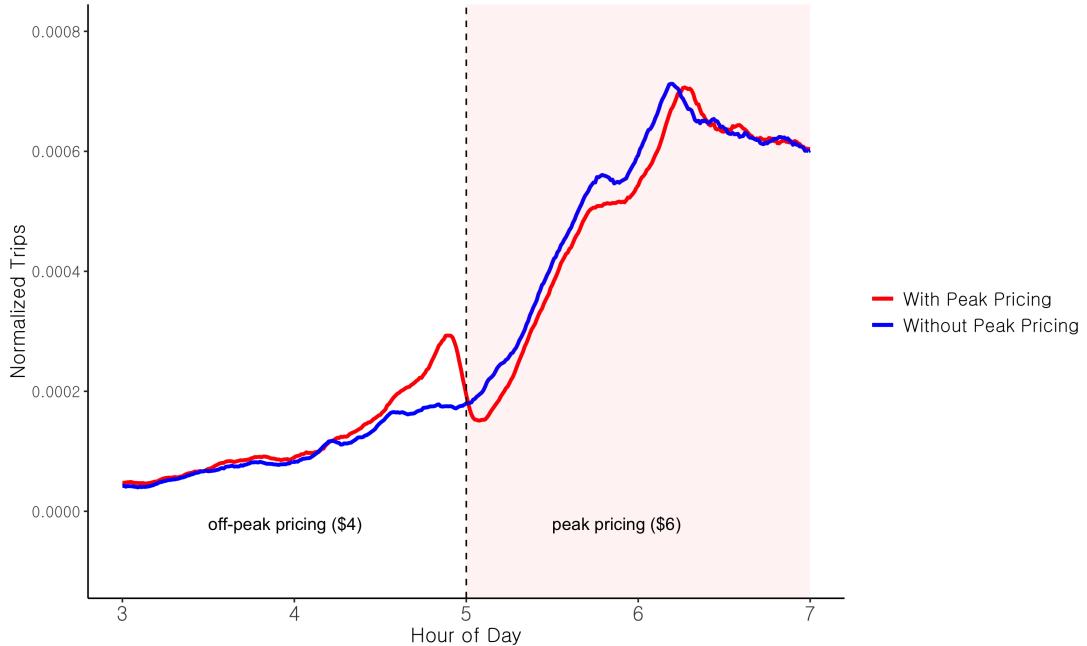


Figure 7: This figure plots the difference in the number of trips in the 6 months before (blue) vs the 6 months after (red) the imposition of peak-hour pricing on the Bay Bridge on July 1, 2010. To facilitate comparison, the number of trips at each time of day is normalized (divided by the total number of daily pre or post-period vehicle trips). The red shaded region demarcates times of day that were subject to peak-hour pricing after July 1, 2010. The vehicle trip counts reflect administrative tolling microdata collected by the Bay Area Toll Authority. Excluded from this graph are trips using the carpool/EV lane, which face a different pricing scheme. Figures 15 and 16 plot bunches for the other price notches (10 a.m., 3 p.m., and 7 p.m.) in the peak-hour pricing scheme on the Bay Bridge.

Table 3 — ESTIMATING SCHEDULING COSTS VIA BUNCHING

Parameter	Estimate (\$\backslash hour)
Time Early (5 a.m. notch)	6.19461
Time Early (3 p.m. notch)	9.74395
Time Late (10 a.m. notch)	15.49765
Time Late (7 p.m. notch)	13.75856

Table 3: This table shows estimates of the costs to drivers of scheduling trips earlier or later than the driver’s ideal trip time (γ_e and γ_l in equation 13). I recover these estimates using equation 23, which relates scheduling costs to the number of additional vehicle trips observed in the period just outside of the peak-hour pricing period on San Francisco’s Bay Bridge. In addition to the number of extra trips, equation 23 reflects scheduling frictions, as well as any time savings that result from drivers adjusting their trips to fall just outside of peak hours (assuming a \$20 value of travel time for Bay-Area travelers, as estimated by [Goldszmidt et al. \(2020\)](#)). The additional bunching mass at price notches is estimated by comparing the number of trips in the neighborhood of the threshold time before vs. after the imposition of peak-hour pricing (see equation 21) using administrative tolling data from the Bay Area Toll Authority. All values are in 2010 dollars.

7.3. Comparisons to Parameter Estimates from the Literature

Several studies from the transportation economics literature provide valuable context for the logit and bunching estimator results presented in this section.

A common heuristic for the value of travel time is 50% of the wage rate, which reflects seminal work by [Lave \(1969\)](#), as well as research collated by [Small \(2012\)](#). According to the 2010 - 2012 California Household Transportation Survey, the median Bay Area household earned roughly \$66,000 per worker, equivalent to \$31.74 per hour. The 50% heuristic therefore implies a median value of travel time of just under \$16. Recent empirical estimate suggest slightly higher travel time: using a field experiment among Lyft riders, [Goldszmidt et al. \(2020\)](#) recover estimates of the value of travel time in San Francisco equal to roughly \$20, or roughly 75% of the 2017 after-tax wage rate (\$17.79 in 2010 dollars).

Estimates of scheduling costs (γ_e and γ_l) are less common in the economics literature. In general, existing studies accord with the canonical analysis by [Small \(1982\)](#), which found that a) it is more costly for drivers to be late than early, b) on a per-hour basis, the cost of being early is lower than the value of travel time, and c) the cost of being late can be higher or lower than the value of travel time depending on the setting. [Kreindler \(2018\)](#), for example, estimates that for drivers in Bangalore, India, early-arrival schedule costs are roughly a quarter of the value of travel time, and late-arrival is more costly than early arrival. In a 2005 choice experiment, [Tseng, Ubbels, and Verhoef \(2005\)](#) find that for drivers in the Netherlands, the cost of early arrival (€4.9/hour) is roughly half of the value of travel time (€9.8/hour), but late arrivals are very costly (€19.7/hour).

The scheduling costs I recover using discrete choice and bunching estimators are qualitatively similar to previous findings: The bunching estimator suggest that drivers prefer being early to being late. Both estimation strategies suggest that (on a per-hour basis) early and late scheduling costs are lower than the value of travel time.

8. Second-Best Optimal Cordon Prices

In this section I use the discrete choice model estimated in Section 7 to calculate optimal cordon prices. I first demonstrate this procedure 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 National Household Transportation Survey) 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 described 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 Francisco’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 8 shows a map of the proposed cordon zone, and Table 4 the proposed tolling schedule.

In the main results presented in this section, I treat as fixed the shape of the cordon and the time periods where prices will be charged. Doing so accords with the setup of the second-best tax problem described in Section 2, where the set of taxable goods is an exogenous constraint. Here the set of taxable goods, J , includes only two goods: morning and evening peak-hour trips that pass through the cordon zone. I present results from expanding the set of taxable goods in Section 8.8.

For simplicity, I also assume that all passenger vehicles will be charged the same price for using the cordon zone. This assumption abstracts from the low-income cordon price exemptions being considered by planning organizations in many cities. In Appendix J, I show that because the majority of commuters would not qualify for this exemption, the changes in welfare, congestion, and pollution that would result from exempting low-income drivers in the San Francisco Bay Area are second-order. As acknowledged in Section 2, the setup of this problem also assumes that policymakers do not weigh marginal utility across income groups. For a full characterization of how policymaker distaste for regressivity impacts corrective taxation, see [Allcott, Lockwood, and Taubinsky \(2019\)](#).

Table 4 — SAN FRANCISCO’S PROPOSED CONGESTION PRICING SCHEME

Time Period	Income Group			
	High	Middle	Low	Very Low
Peak Hours (6-10 a.m., 3-7 p.m.)	6.50	4.33	2.17	Free
Off-Peak Hours	Free	Free	Free	Free

Table 4: San Francisco’s proposed cordon pricing scheme as of September 1, 2021. Trips that enter the cordon (see Figure 8) 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 without dependants, for example, the annual income thresholds for high, middle, low, and very low income are \$150,000, \$116,000, \$66,000 and \$46,000, respectively.

FIGURE 8 — SAN FRANCISCO’S PROPOSED CONGESTION PRICING ZONE

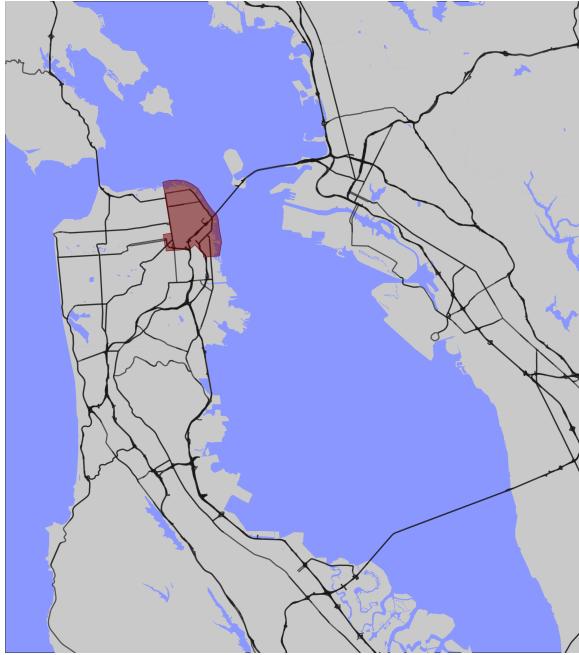


Figure 8: San Francisco’s proposed cordon pricing scheme as of September 1, 2021. Trips that enter the cordon would be charged during peak hours according to the pricing scheme outlined in Table 4. Tolls will be levied using electronic tag readers mounted on gantries that span roadways on the border of the cordon region.

8.2. Personal Vehicle Trips in the San Francisco Area

The National Household Transportation Survey (NHTS) is a survey of US individual travel habits administered by the Federal Highway Administration. Participants in this survey are recruited via mail; survey responses are incentivized by small (\$5 to \$20) rewards, and can be completed through mail-back forms or online. The 2017 NHTS garnered responses from 381,975 individuals, each of whom filled out “Travel Diaries” that detailed their travel habits during one randomly selected 24-hour period. In addition to information about the attributes of the trip taken, the NHTS also collects demographic information about surveyed persons and their households.

I use the 2017 NHTS California Add-On¹⁰ to build a representative dataset of trips that cross San Francisco’s proposed cordon zone.¹¹ Each *trip* in this dataset consists of a start location (zip code or Census Block), an end 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. The resulting dataset contains 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 the left pane of Figure 9.

To predict substitution in time and space under San Francisco’s cordon, I construct a set of alternatives for each trip. For every cordon trip in the NHTS, I construct alternative departure times at 12-minute intervals throughout the day. Using HERE Technology’s *Routes* API, I can assign travel times to each of these alternative trips by varying the departure time. 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

¹⁰The NHTS Add-On program allows States and Cities to pay the Federal Highway Administration to administer additional surveys in their region for the purposes of developing a more comprehensive sample.

¹¹Note that the FasTrak microdata used in Section 7 are ill-suited for this task because many of the trips that cross San Francisco’s proposed cordon do not use any bridge (e.g., trips with two termini within the city of San Francisco).

the cordon zone. I then calculate travel times for this non-cordon route for each 12-minute interval throughout an average traffic day. These detour routes are plotted in the right pane of Figure 9.

The result of this data collation is a set of trip endpoints for the San Francisco area, where drivers can choose over $route \in \{\text{cordon, non-cordon}\}$ and $time\ of\ day \in \{4.0, 4.2, \dots, 22.0\}$, as well as a generic outside option. This choice set allows me to predict how drivers would choose between options based on the attributes (travel time, time early, time late, and toll price) specified by the discrete choice model estimated in Section 3.

FIGURE 9 — CORDON AND NON-CORDON ROUTES FOR BAY-AREA TRIPS

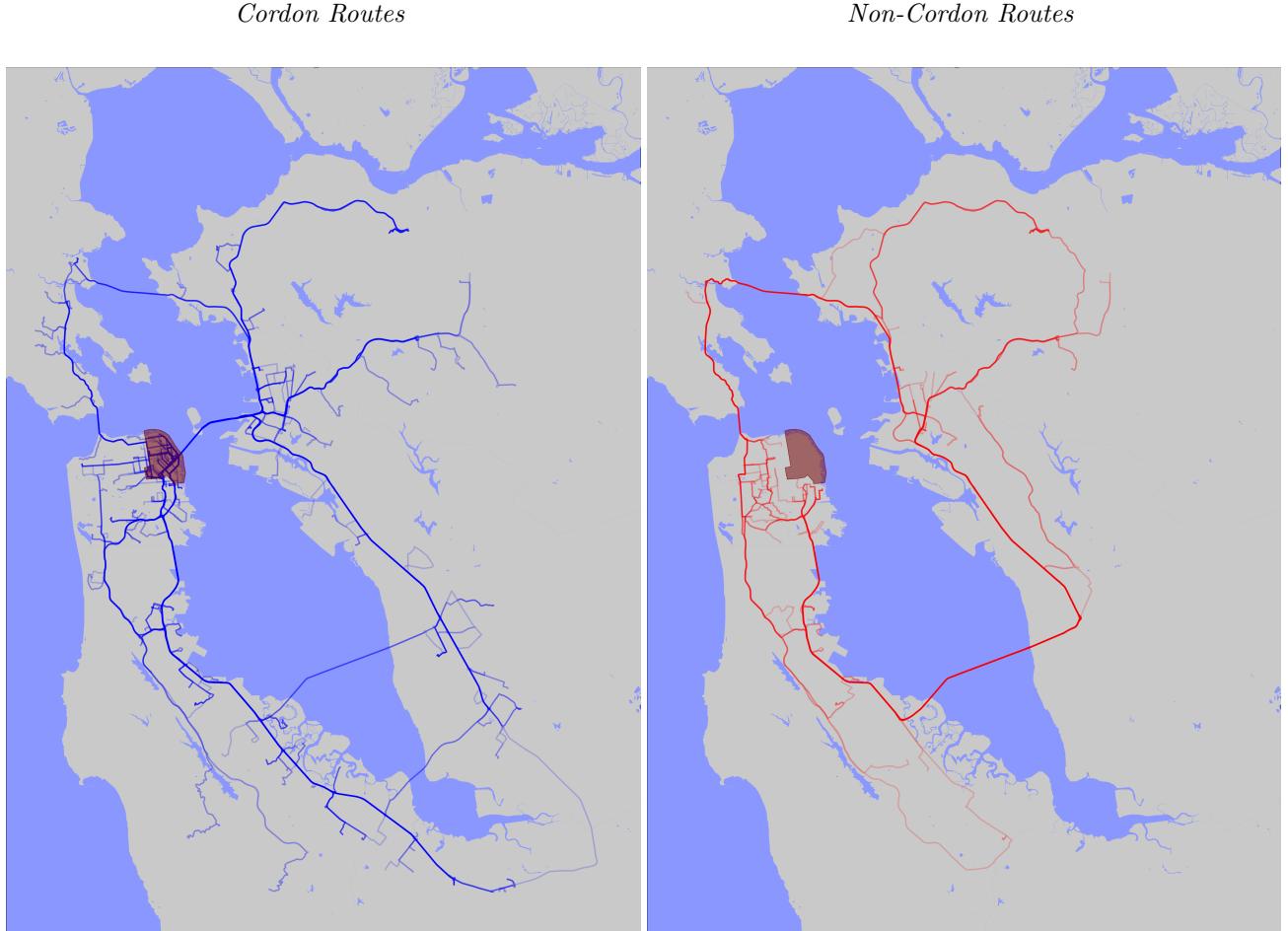


Figure 9: This figure plots cordon and non-cordon routes constructed from the 2017 National Household Transportation Survey (NHTS) California Add-On. The left pane plots 1,891 trips that cross San Francisco’s proposed congestion zone, according to suggested routes generated with the HERE Technology’s *Routes* API. The right pane plots detour routes for the subset of these 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 12-minute time of day interval, as well as a non-cordon route (the right pane), if such a detour exists, for every 12-minute time of day interval. The choice sets of all drivers also include a generic outside option.

8.3. Trip-Level Externalities

For each trip described above (trips in the NHTS with suggested routes that pass through the cordon, as well as alternative trips in space and time), I assign traffic and pollution externalities in a manner similar to the process described in Section 5. The detail of the NHTS survey data, however, allows for more precise estimation

of both congestion and pollution externalities relative to trips observed in the FasTrak tolling data.

As shown in Figure 5, emissions vary by vehicle attributes as well as travel speed. The NHTS includes information about the vehicle used on each trip, including the vehicle vintage, make, and fuel type (gasoline, diesel, EV, or hybrid). Using the travel time and distance information for each trip returned by the HERE Routes API, I 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 10.

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. Following the procedure used to assign externalities to FasTrak trips, I rely on a comprehensive network of traffic sensors on roadways in the Bay Area to estimate the traffic density along each route at different times of day. Concretely, this requires first identifying sensors along the trip’s route, 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, multiplied by the length of the trip. I plot the trip-level externalities for the 1,891 NHTS trips that cross the proposed cordon in Figure 10.

FIGURE 10 — EXTERNAL COSTS FOR TRIPS CROSSING SAN FRANCISCO’S PROPOSED CORDON

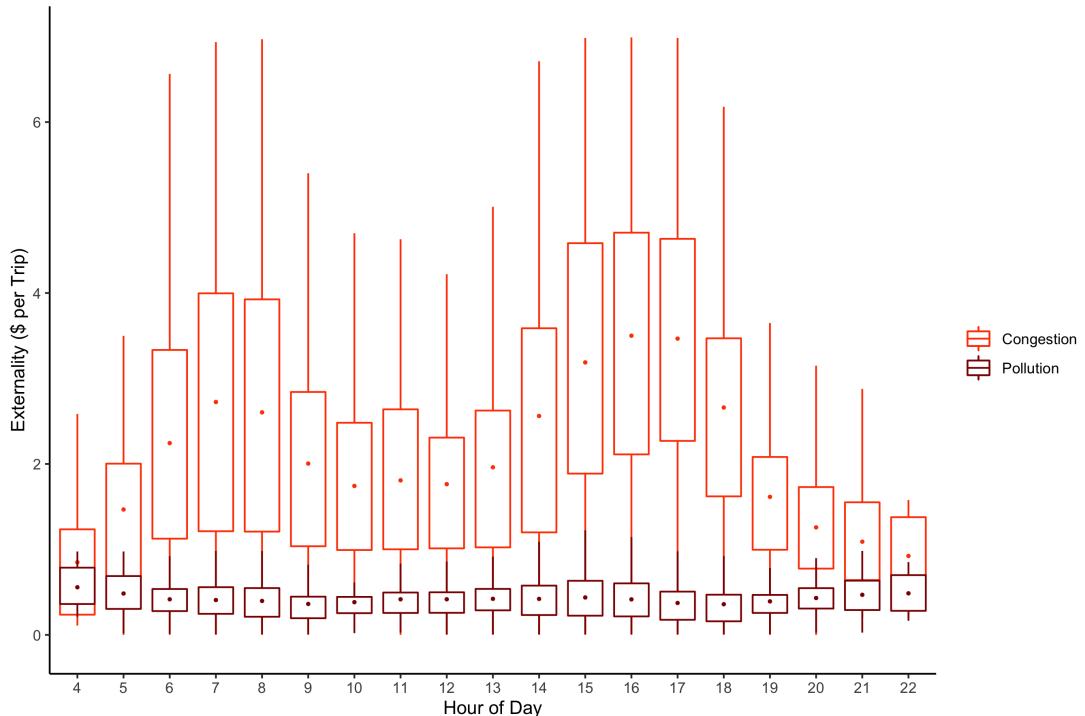


Figure 10: This Figure plots pollution (brown) and congestion (orange) externalities by hour for trips in the 2017 National Household Transportation Survey (NHTS) with suggested routes that pass through San Francisco’s proposed cordon zone. The mean externality within any given hour is represented by a dot; the box spans the 25th to 75th externality percentile, and the bars span the 5th to 95th externality percentile. Trip routes reflect the suggested directions 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 \(2020\)](#) and an average value of travel time of \$20, as per [Goldszmidt et al. \(2020\)](#). 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.

8.4. Substitution Between Trips

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 equation 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 $\in \{\text{cordon, non-cordon}\}$ at a specific hour of day $\in \{4.0, 4.2, \dots, 22.0\}$) 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 of 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 that circumvent the cordon (call these trips z). The leakage shares 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 estimated in Section 7.¹² 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 for individual i is:

$$\frac{\partial_{\{h^j, r^k\}}}{\partial p_{\{h^l, r^m\}}} = \int \beta_i s_{\{h^j, r^k\}}(\theta) s_{\{h^l, r^m\}}(\theta) f(\theta) d\theta \quad (24)$$

where θ is the joint distribution of random coefficients in the mixed logit model, β_i is the logit parameter governing price responsiveness for individual i , and $s_{\{h^j, r^k\}}(\theta)$ is the share predicted trips that take route k at time j under the random coefficient vector θ .

8.5. Optimal Prices

Figure 11 plots three lines relevant 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 2. Finally, the red line plots the second-best optimal prices for San Francisco's proposed cordon when tolling is restricted to morning and evening peak hours (6-10 a.m. and 3-7 p.m., respectively). The second-best optimal scheme charges \$2.20 during morning peak hours, and \$2.85 during evening peak hours. These second-best optimal prices are calculated using equation 10, and take into account both the correlation between externalities and elasticities, as well as the substitution to unpriced alternatives in time and space.

¹²Note 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 11 — SECOND-BEST OPTIMAL ROAD PRICES

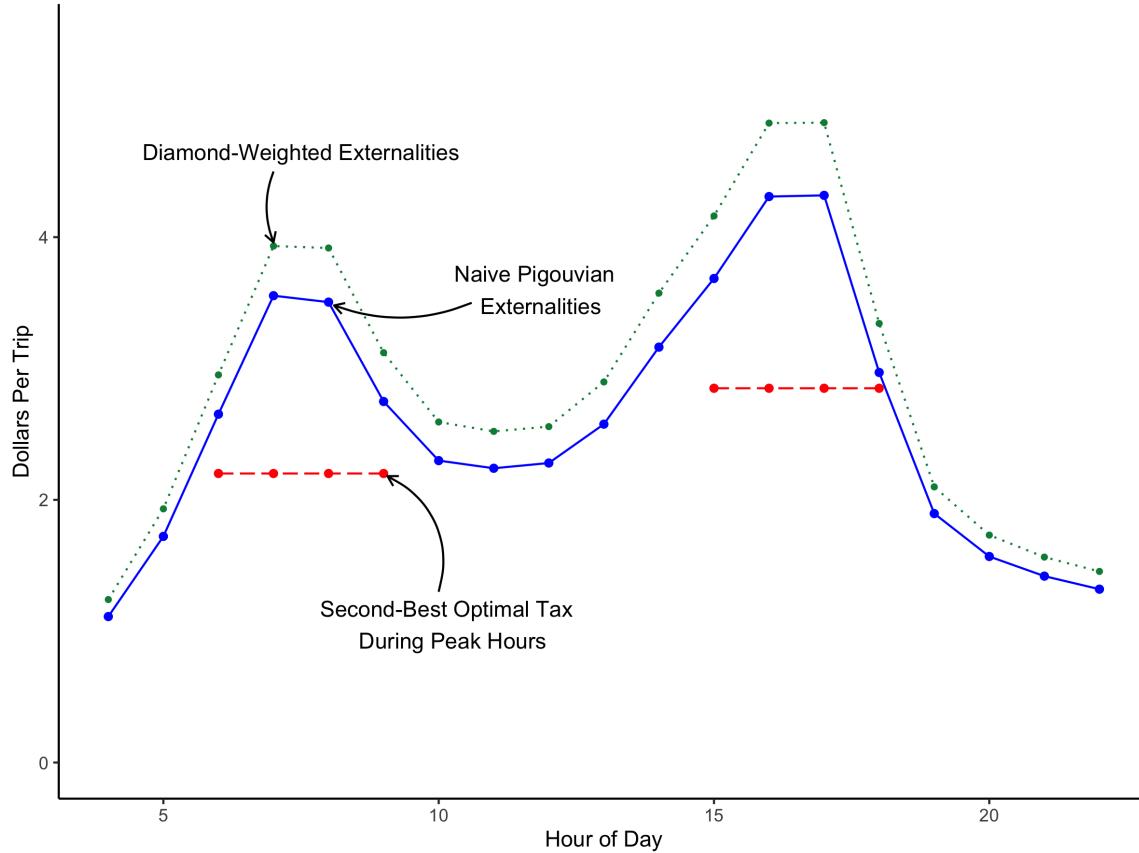


Figure 11: This figure plots three prices relevant for understanding optimal second-best cordon tolls. The blue (solid) line plots the average externality (pollution and congestion) for trips that cross San Francisco’s cordon by hour of day, estimated using data from the 2017 NHTS (see Section 8.3). 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 (6-10 a.m. and 3-7 p.m., respectively). These second-best optimal prices are calculated using equation 10, and takes into account both the correlation between externalities and elasticities (“Diamond weights”), as well as the substitution (leakage) to unpriced alternative trips in time or space.

The results plotted in Figure 11 reflect social damages calculated using driving conditions that exist in the current, untaxed equilibrium. Consistent with the literature on externality taxation, the second-best tax formula presented in Section 8 phrases optimal taxes as a function of externalities at the optimum. As shown in figures 4 and 25, the marginal damages associated with driving are non-constant in traffic density/speed, meaning that in general, damages at the taxed equilibrium will be different (lower) than those observed in the untaxed equilibrium. Whether or not the difference between marginal damages calculated at versus away from the optimum is a first-order concern depends on the slope of the marginal damages function and the responsiveness of drivers to taxation. In Appendix F, I use simulations where I iteratively calculate taxes and traffic density to bound the second-best optimal cordon prices in San Francisco. The fixed point from this exercise constitutes a lower bound because it ignores “induced demand”, which will tend to attenuate the difference in traffic conditions between taxed and untaxed equilibria (Duranton and Turner, 2011). I recover lower bounds of \$1.60 and \$1.80 for the morning and evening peak-hours, respectively.

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 under 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 \$2.20 during morning peak hours, and \$2.85 during evening peak hours (see Figure 11).

I plot outcomes from these simulations in Figures 17 through 19, 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 total trips, congestion, and pollution just outside peak-hour pricing periods.

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 a first-best policy where all vehicle trips (all times of day; cordon and non-cordon) are charged based on the social damages they generate. The four outcomes of interest are total externalities (pollution and congestion), congestion alone, pollution alone, and total welfare (the utility of all drivers, in dollars, less total externalities). The results in this table reflect 600,000 simulated choices by drivers in the NHTS dataset constructed above — 600,000 is roughly the number of weekday trips that pass through San Francisco’s proposed cordon, according to the San Francisco County Transportation Authority. The choice probabilities for different alternatives (cordón vs. non-cordón trips at different times of day, and a generic outside option) were generated by applying the mixed logit model shown in Table 2 to the NHTS driver choice sets constructed in Section 8.

8.7. Cordon Pricing in New York and Los Angeles

In addition to San Francisco, city governments in New York and Los Angeles are also considering implementing cordon pricing zones (mapped in Figure 12). In this section, I calculate optimal peak-hour 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 12 — PROPOSED CORDONS IN NEW YORK AND LOS ANGELES

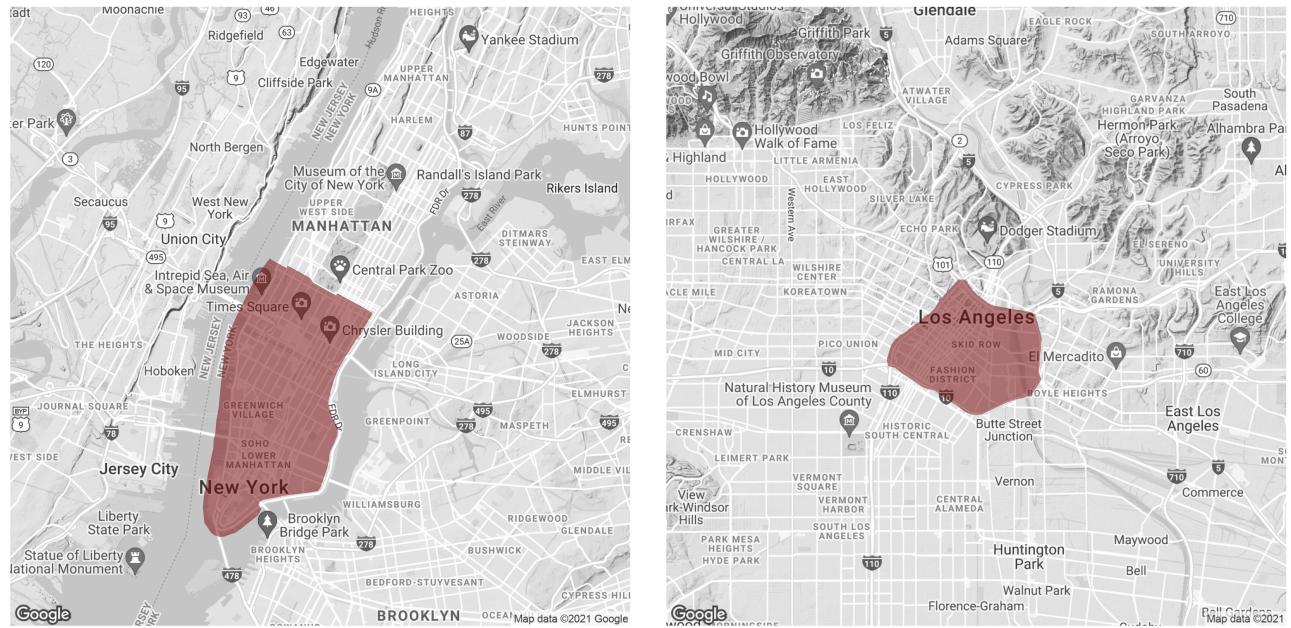


Figure 12: Proposed cordon pricing schemes in New York and Los Angeles. All proposals are as of August, 2021. The New York congestion map is courtesy of the *Regional Plan Association*; The Los Angeles map is courtesy of the *LA Metro*.

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 data¹³ and Here Technology’s *Routes* API to identify trips where the fastest route passes through the city’s proposed cordon. Second, I use vehicle attributes and travel speed to assign pollution externalities, and use traffic density data¹⁴ from city roads to assign congestion externalities to those trips. Third, I 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 driving demand primitives (price responsiveness, β , scheduling costs, γ_e and γ_l , and the value of travel time, α) that are used to calculate substitution parameters, as well as city-specific correlations between externalities and price responsiveness (Diamond weights). Absent such experiments, I calculate optimal cordon prices and welfare outcomes in New York and Los Angeles using the driving demand primitives and Diamond weights estimated in San Francisco (see Table 2). These results are reported in Tables 6 and 7. In Appendix I, I use questions from the 2017 NHTS to examine the external validity of the model estimated in San Francisco. 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. These proxies for demand primitives are broadly similar across the three cities I examine in this paper.

¹³The 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 come from the 2018 NY Citywide Mobility Survey.

¹⁴Traffic density data for Los Angeles is publicly available through PeMS. Traffic density for NY is courtesy of the NYSDOT Traffic Monitoring Section.

Table 6 — COMPARING SECOND-BEST CORDON PRICES TO SOCIAL DAMAGES

Period	Value (\$)		
	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 Social Damages, AM Peak (6-10)	3.115	4.877	8.186
Average Social 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.

Table 7 — CONGESTION, POLLUTION, AND WELFARE EFFECTS OF PEAK-HOUR CORDON PRICING

Outcome	Performance Relative to the First-Best (%)		
	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 a first-best policy where all vehicle trips (all times of day; cordon and non-cordon) are charged based on the social damages they generate. “Peak hours” are defined as 6-10 a.m. and 3-7 p.m.; 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 results in this table reflect the simulated choices of 600,000 (SF and LA) to 1 million (NY) drivers. The choice probabilities for different alternatives (cordon vs. non-cordon trips at different times of day, and a generic outside option) were generated by applying the mixed logit model shown in Table 2 to the driver choice sets constructed using transportation survey data (see Section 8).

8.8. Hourly Cordon Pricing

Tables 5, 6, and 7 describe results where the policymaker is restricted to only price cordon trips during peak hours, as is proposed by the San Francisco County Traffic Authority. In this section I relax this constraint, allowing the policymaker to set a fixed hourly toll schedule during normal commuting times. In the notation of the second-best tax model outlined in Section 2, The set J now includes 14 taxable “goods,” where each good covers all cordon trips for a given hour of day $\in \{6, 7, \dots, 19\}$.

Tables 8 and 9 display estimates of welfare outcomes under second-best tax with hourly cordon pricing versus a first-best policy where every trip is charged according to the social damages associated with that trip. Relaxing this constraint leads to significant welfare improvements in three cities relative to a peak-hour scheme — the welfare gains are roughly twice as large in San Francisco and Los Angeles, and 16% higher in New York

City. In each of the three cities, however, this more flexible policy fails to achieve more than half of the welfare gains that would be realized under a first-best policy, due to a combination of externality leakage, and uniform prices charged to heterogeneous users of the cordon zone within each hour.

Table 8 — CONGESTION, POLLUTION, AND WELFARE EFFECTS OF HOURLY CORDON PRICING

Outcome	Performance Relative to the First-Best (%)		
	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	49.706

Table 8: This table summarizes the performance of second-best optimal cordon pricing schemes when policymakers are allowed to set a fixed schedule of tolls between 6 a.m. and 7 p.m., relative to a first-best policy where every trip is charged according to its social damages. 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 results in this table reflect simulated choices using the mixed logit model shown in Table 2. The number of simulated trips reflects the number of daily trips in the cordon regions proposed in each city (600,000 in San Francisco and Los Angeles, and 1,040,000 in New York), as estimated by the planning authorities responsible for designing the cordon in each respective city. 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

Policy	Welfare Gain Relative to the Status Quo (\$ Million)		
	San Francisco	Los Angeles	New York
First-Best	852	1,100	1,477
Second-Best (Peak Only)	246	166	630
Second-Best (Fixed Hourly)	412	323	715

Table 9: This table displays back of the envelope calculations for the annual welfare gains under three road pricing policies: 1) The first-best 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) second-best-optimal time-of-day prices, which are allowed to vary by hour according to a fixed schedule between 6 a.m. and 7 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. The figures in this table reflect simulated choices using the mixed logit model shown in Table 2. The number of simulated trips reflect the number of daily trips in the cordon regions proposed in each city (600,000 in San Francisco and Los Angeles, and 1,040,000 in New York), as estimated by the planning authorities responsible for designing the cordon in each respective city.

9. Discussion

Cordon prices differ from a first-best driving tax in two important ways: 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 (Table 2) implies 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 11 show that the leakage effect strongly dominates in the case of San Francisco’s cordon zone: Optimal prices are \$2.20 for the morning peak period and \$2.85 for the evening peak period — roughly half of the average social cost for trips that pass through the cordon at those times.

My findings suggest that if the primitive determinants of driver decisions (price responsiveness, value of travel time, schedule flexibility) are similar across cities, then optimal cordon prices are also below the average of social damages generated by downtown trips in New York and Los Angeles. Table 6 shows that in New York, for example, 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 morning and evening peak prices are \$3.30 and \$4.53, compared to average social damages of \$4.87 and \$5.72.

Cordon zones charging the second-best prices described in Table 6 would generate significant welfare gains for commuters and city residents in all three cities. The benefits from optimal peak-hour cordon prices range from \$246 million annually in San Francisco to \$426 million annually in New York (Table 9). 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.

Despite these gains, the results in Section 8 suggest that the blunt nature of cordon pricing limits their effectiveness relative to an ideal policy. 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 Figure 10.

What adjustments could improve the performance of the proposed cordon zones in the United States? Relative to a peak-only tolling scheme, allowing policymakers to set a fixed schedule of prices that vary by time of day (Table 8) provides sizeable welfare gains: \$166 million in San Francisco, \$151 million in Los Angeles, and \$85 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.

The theory provided in Section 2 suggests that policymakers may want to set boundaries to preempt spatial leakage. Depending on the idiosyncratic geography of a city, an optimal cordon zone may include outlying or relatively uncongested routes that provide close substitutes for congested central roads. Expanding cordon zones, however, comes at a cost; regardless of the design of the tolling system at the boundaries, trips that remain entirely inside the cordon are not priced. Expanding the cordon too far may therefore undermine the policy’s overall coverage. A full characterization of this tradeoff is beyond the scope of this paper, but may prove an interesting question for future research.

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 of 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.
- 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.
- Duranton, Gilles and Matthew A Turner. 2011. “The fundamental law of road congestion: Evidence from US cities.” *American Economic Review* 101 (6):2616–52.

- Finkelstein, Amy. 2009. “E-ztax: Tax salience and tax rates.” *The Quarterly Journal of Economics* 124 (3):969–1010.
- Foreman, Kate. 2016. “Crossing the bridge: The effects of time-varying tolls on curbing congestion.” *Transportation Research Part A: Policy and Practice* 92:76–94.
- 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.
- Goldszmidt, Ariel, John A List, Robert D Metcalfe, Ian Muir, V Kerry Smith, and Jenny Wang. 2020. “The Value of Time in the United States: Estimates from Nationwide Natural Field Experiments.” Tech. rep., National Bureau of Economic Research.
- 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, Jinyok, 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.
- Herrnstadt, Evan, Ian WH Parry, and Juha Siikamäki. 2015. “Do alcohol taxes in Europe and the US rightly correct for externalities?” *International Tax and Public Finance* 22 (1):73–101.
- 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.
- 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.
- 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* .
- Lave, Charles A. 1969. “A behavioral approach to modal split forecasting.” *Transportation Research UK* 3 (4).
- 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.
- Regional Plan Association. 2021. “Congestion Pricing in New York City.” Tech. rep., Regional Plan Association.
- 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. 1982. “The scheduling of consumer activities: work trips.” *The American Economic Review* 72 (3):467–479.
- . 2012. “Valuation of travel time.” *Economics of Transportation* 1 (1-2):2–14.
- 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.
- Tseng, Yin Yen, Barry Ubbels, and Erik Verhoef. 2005. “Value of time, schedule delay and reliability.” In *ERSA conference papers*.
- 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, \dots, h_M) and a numeraire, z . Each non-numeraire good has an associated externality, ϕ_m . A policymaker can choose tax levels for goods $j \in \{1, \dots, J\}$ where $J < M$. I assume goods $k \notin \{1, \dots, J\}$ are un- or under-taxed.

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

$$\max\{U(h_1, \dots, h_M) + z\} \quad s.t. \quad (25)$$

$$(p_1 + \tau_1)h_1 + (p_J + \tau_J)h_J + p_{J+1}h_{J+1} + \dots + p_Mh_M + z \leq I \quad (26)$$

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

$$U_j = \lambda(p_j + \tau_j) \quad \forall \quad j \in \{1, \dots, J\} \quad (27)$$

$$U_k = \lambda(p_k) \quad \forall \quad k \notin \{1, \dots, J\} \quad (28)$$

$$\lambda = 1 \quad (29)$$

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 $\tau_1 \dots \tau_J$, which are applied to the taxable goods $j \in \{1, \dots, J\}$.

$$\begin{aligned} \max\{U(h_1, \dots, h_M) + z - \sum_1^M \phi_m h_m\} \quad & st. \\ p_1 h_1 + \dots + p_N h_N + z \leq I \end{aligned} \quad (30)$$

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

$$0 = \frac{\partial h_j}{\partial p_j} [U_j - \phi_j - p_j] + \sum_{k \neq j}^M \frac{\partial h_k}{\partial p_j} [U_k - \phi_k - p_k] \quad (31)$$

Plugging in the consumer's first order conditions and solving for $\tau_m \dots$

$$0 = \frac{\partial h_j}{\partial p_j} [\tau_j - \phi_j] + \sum_{k \neq j}^J \frac{\partial h_k}{\partial p_j} [\tau_k - \phi_k] + \sum_{l=J+1}^M \frac{\partial h_l}{\partial p_j} [\phi_l] \quad (32)$$

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

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:

$$\tau_j + \frac{1}{\frac{\partial h_j}{\partial p_j}} \left(\sum_{k \neq j}^J \frac{\partial h_k}{\partial p_j} \tau_k \right) = \phi_j + \frac{1}{\frac{\partial h_j}{\partial p_j}} \sum_{l=1}^M \frac{\partial h_l}{\partial p_j} \phi_l \quad (34)$$

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

$$a_1^j \tau_1 + \dots + a_k^j \tau_k + \dots + a_J^j \tau_J = b_j \quad \forall j \in \{1, \dots, J\} \quad (35)$$

Where a_k^j and b_m are defined as:

$$a_k^j = \frac{\frac{\partial h_k}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \quad (36)$$

$$\beta_j = \phi_j + \sum_{l=1}^M \frac{\frac{\partial h_l}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \phi_l \quad (37)$$

The a and β terms have an intuitive interpretation. a_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 from 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:

$$\begin{bmatrix} a_1^1 & \dots & a_1^J \\ \vdots & \ddots & \vdots \\ a_1^J & \dots & a_J^J \end{bmatrix} \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_J \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_J \end{bmatrix} \quad (38)$$

$$\mathbf{A}\boldsymbol{\tau} = \mathbf{b} \quad (39)$$

$$\boldsymbol{\tau} = \mathbf{A}^{-1} \mathbf{b} \quad (40)$$

A.2. Heterogeneity and Leakage

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): $U_i(h_1^1, \dots, h_1^M, \dots, h_i^1, \dots, h_i^M, \dots, h_N^1, \dots, h_N^M) + z_i$.

As in section 2.3, a policymaker can choose tax levels for goods $j \in \{1, \dots, J\}$ where $J < M$. I assume goods $k \notin \{1, \dots, J\}$ 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, \dots, h_i^M) and their consumption of the numeraire good z_i .

$$\begin{aligned} & \max\{U_i(h_1^1, \dots, h_1^M, \dots, h_i^1, \dots, h_i^M, \dots, h_N^1, \dots, h_N^M) + z_i\} \text{ st.} \\ & (p^1 + \tau^1)h_i^1 + (p^J + \tau^J)h_i^J + p^{J+1}h_i^{J+1} + \dots + p^M h_i^M + z_i \leq \mu_i \end{aligned} \quad (41)$$

The first-order conditions for this problem are:

$$\begin{aligned} \frac{\partial U_i}{\partial h_i^j} &= \lambda(p^j + \tau^j) \quad \forall j \in \{1, \dots, J\} \\ \frac{\partial U_i}{\partial h_i^k} &= \lambda(p^k) \quad \forall k \notin \{1, \dots, J\} \\ \lambda &= 1 \end{aligned} \quad (42)$$

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 $\tau^1 \dots \tau^J$, which are applied to the taxable goods $j \in [1, J]$.

$$\begin{aligned} & \max\left\{\sum_i^N(U_i(h_1^1, \dots, h_1^M, \dots, h_i^1, \dots, h_i^M, \dots, h_N^1, \dots, h_N^M) + z_i)\right. \\ & \text{st.} \quad \left.(p^1)\sum_i^N h_i^1 + \dots + (p^J)\sum_i^N h_i^J + (p^{J+1})\sum_i^N h_i^{J+1} + \dots + (p^M)\sum_i^N h_i^M + \sum_i^N z_i \leq \sum_i^N \mu_i\right. \end{aligned} \quad (43)$$

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

$$0 = \sum_{i=1}^N \frac{\partial U_i}{\partial h_i^j} \frac{\partial h_i^j}{\partial p_j} + \sum_{i=1}^N \sum_{g \neq i}^N \frac{\partial U_i}{\partial h_g^j} \frac{\partial h_g^j}{\partial p_j} + \dots + \frac{\partial U_i}{\partial h_g^M} \frac{\partial h_g^M}{\partial p_j} - p^1 \sum_i \frac{\partial h_i^1}{\partial p_j} - \dots - p^M \sum_i \frac{\partial h_i^M}{\partial p_j} \quad (44)$$

Plugging in the consumer's first order conditions and solving for $\tau_j \dots$

$$\tau_j = \frac{\sum_{i=1}^N \sum_g^N \left(\frac{\partial U_i}{\partial h_g^1} \frac{\partial h_g^1}{\partial p_j} + \dots + \frac{\partial U_i}{\partial h_g^M} \frac{\partial h_g^M}{\partial p_j} \right)}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} + \frac{\sum_{k \neq j}^J \frac{\partial h_i^k}{\partial p_j} \tau_k}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} \quad (45)$$

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. Additional Figures and Tables

FIGURE 13 — TRAFFIC SENSORS IN THE BAY AREA



Figure 13: This figure plots traffic sensors from the Caltrans Performance Measurement System (PeMS). Each sensor reports hourly vehicle count and speed data that are converted to traffic density (vehicles/lane/mile) using the fundamental equation of traffic flow. These traffic density readings are then used to assign congestion externalities to vehicle trips based on route and time of day, as described in sections 5 and 8.

[This space is left intentionally blank]

FIGURE 14 — MIXED LOGIT RESULTS

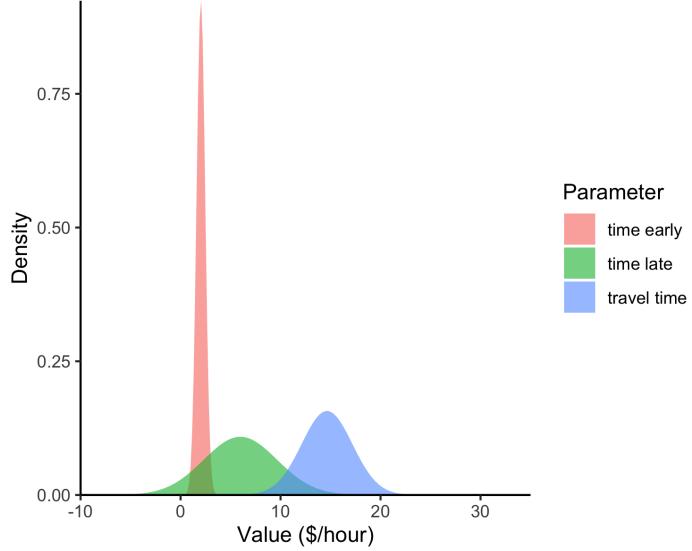


Figure 14: This figure displays the mean and standard deviation from a random coefficients (“mixed”) logit regression used to estimate the parameters in Equation 13. The data used in this regression are the FasTrak tolling microdata described in section 5: A panel of 32,104 individuals and 1,078,044 bridge crossings between June 15, 2010 and July 15, 2010, excluding weekends and holidays. The dependent variable is an indicator variable for 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 ; *time late* is defined analogously. *Price* is the toll that driver i would incur by traveling via route r at hour h . The mean and standard deviation of all time-related variables have been normalized relative to the coefficient on the *price* variable.

Table 10 — MIXED LOGIT RESULTS

variable	mean	sd
Time Early (\$/hr)	2.033	0.175
Time Late (\$/hr)	5.984	1.032
Travel Time (\$/hr)	14.647	1.491
Price	1.000	0.135

Table 10: This table 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 the FasTrak tolling microdata described in section 5: A panel of 32,104 individuals and 1,078,044 bridge crossings between June 15, 2010 and July 15, 2010, excluding weekends and holidays. The dependent variable is an indicator variable for 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 ; *time late* is defined analogously. *Price* is the toll that driver i would incur by traveling via route r at hour h . The mean and standard deviation of all time-related variables have been normalized relative to the coefficient on the *price* variable. The mean and standard deviation of all time-related variables have been normalized relative to the coefficient on the *price* variable. All values are in 2010 dollars.

FIGURE 15 — BUNCHING AT PRICE NOTCHES

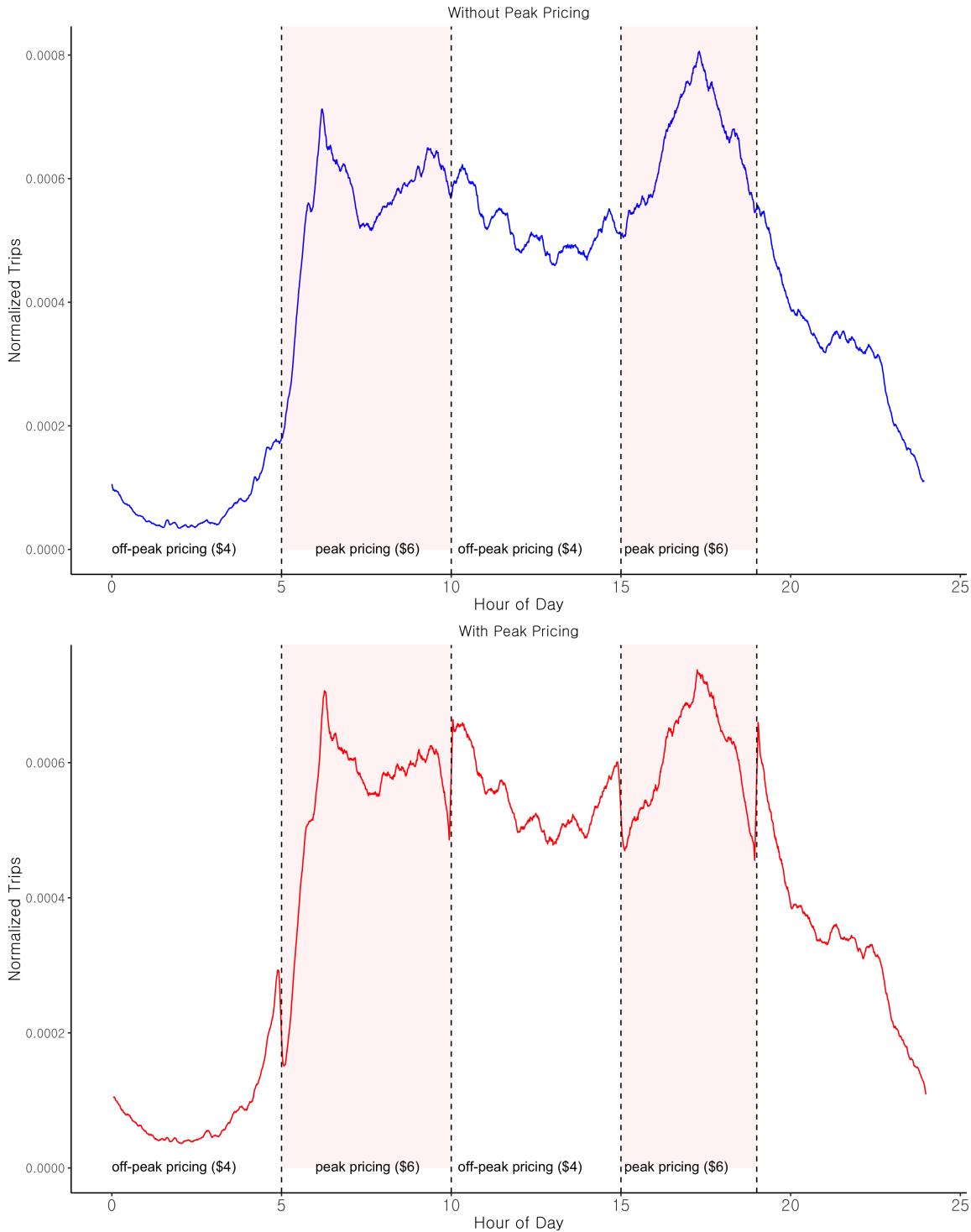


Figure 15: This figure plots the density (the share of total daily crossings) of passenger vehicle trips crossing San Francisco's *Bay Bridge* in the 6 months before (blue) and 6 months after (red) the imposition of peak hour pricing on July 1, 2010. This plot excludes trips that use the carpool lane, as well as eligible electric vehicles, each of which faced a different pricing scheme. The red shaded regions demarcate times of day that were subject to peak-hour pricing after July 1, 2010.

FIGURE 16 — DETAIL OF BUNCHING AT PRICE NOTCHES

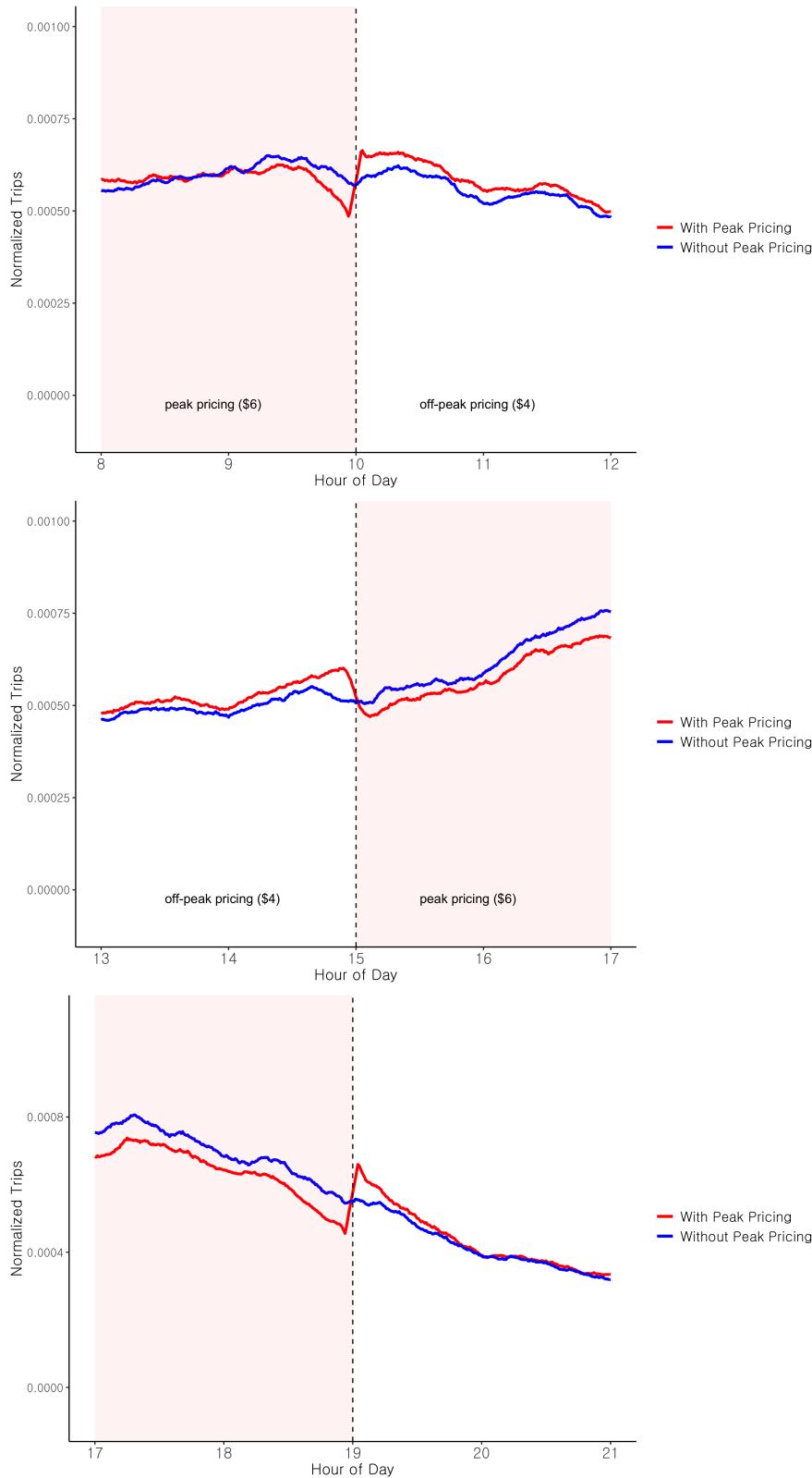


Figure 16: San Francisco's *Bay Bridge* imposed peak hour pricing on July 1, 2010 (see Section 4). This figure plots the density of passenger vehicle trips crossing the Bay Bridge in the 6 months before (blue) and 6 months after (red) the imposition of peak hour pricing for the 10 a.m., 3 p.m., and 7 p.m. price notches. This plot excludes trips that use the carpool lane, as well as eligible electric vehicles, each of which faced a different pricing scheme. The red shaded regions demarcate times of day that were subject to peak-hour pricing after July 1, 2010.

FIGURE 17 — SIMULATED TRAVEL CHOICES UNDER PEAK-HOUR CORDON PRICING IN SAN FRANCISCO

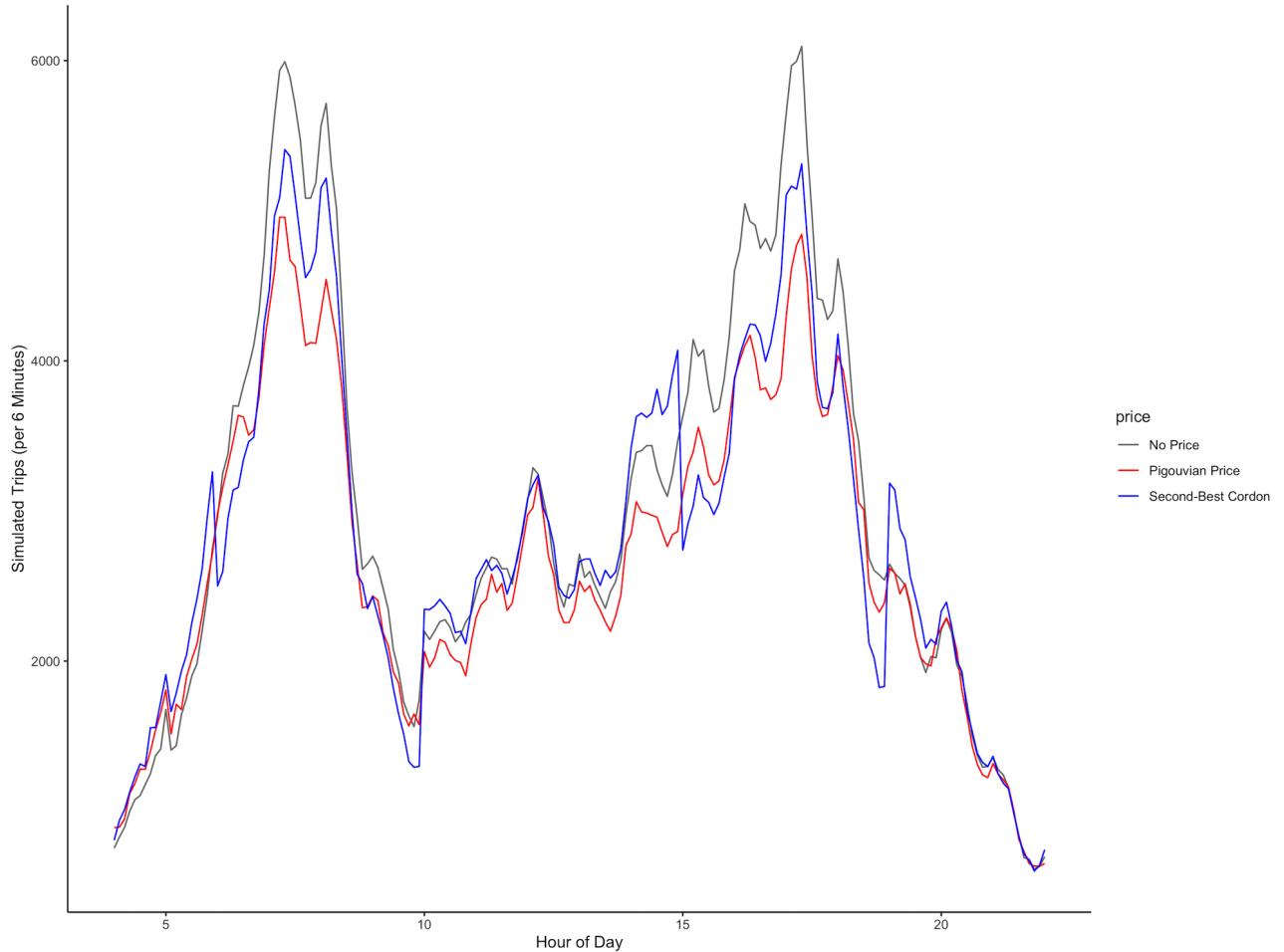


Figure 17: In this figure I plot the number of trips that pass through or near the cordon under three simulations using the mixed logit model estimated in Table 2 of Section 7 together with the NHTS trip dataset described in Section 8. In each scenario, I predict 600,000 choices — roughly daily total 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 first-best scheme where every trip a driver could choose (including non-cordon trips) would be priced according to its marginal pollution and congestion externalities. The red line plots trips under the second-best optimal peak-hour cordon price from Figure 11. Note that all lines include both trips that cross through the cordon, and “detour” trips that circumvent the cordon.

FIGURE 18 — SIMULATED CONGESTION UNDER PEAK-HOUR CORDON PRICING IN SAN FRANCISCO

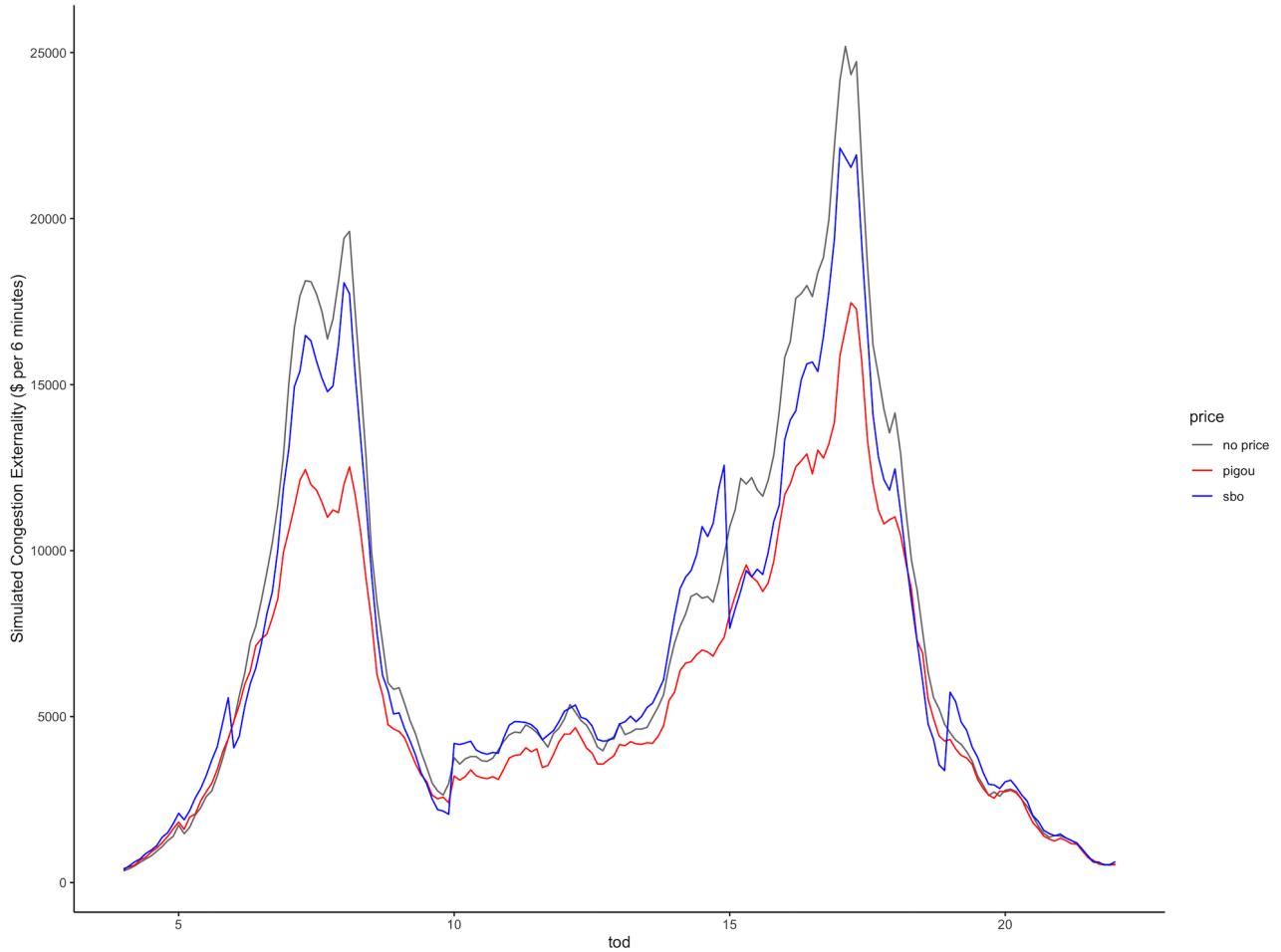


Figure 18: In this figure I plot the total congestion externalities under three simulations using the mixed logit model estimated in Table 2 of Section 7 together with the NHTS trip dataset described in Section 8. In each scenario, I predict 600,000 choices — roughly daily total of vehicle trips that 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 first-best policy where every trip a driver could choose (including non-cordon trips) would be priced according to its marginal pollution and congestion externalities. The red line plots sum of congestion externalities under the second-best optimal peak-hour cordon price from Figure 11. Note that all lines include congestion from trips that cross through the cordon, as well as “detour” trips that circumvent the cordon.

FIGURE 19 — SIMULATED POLLUTION UNDER PEAK-HOUR CORDON PRICING IN SAN FRANCISCO

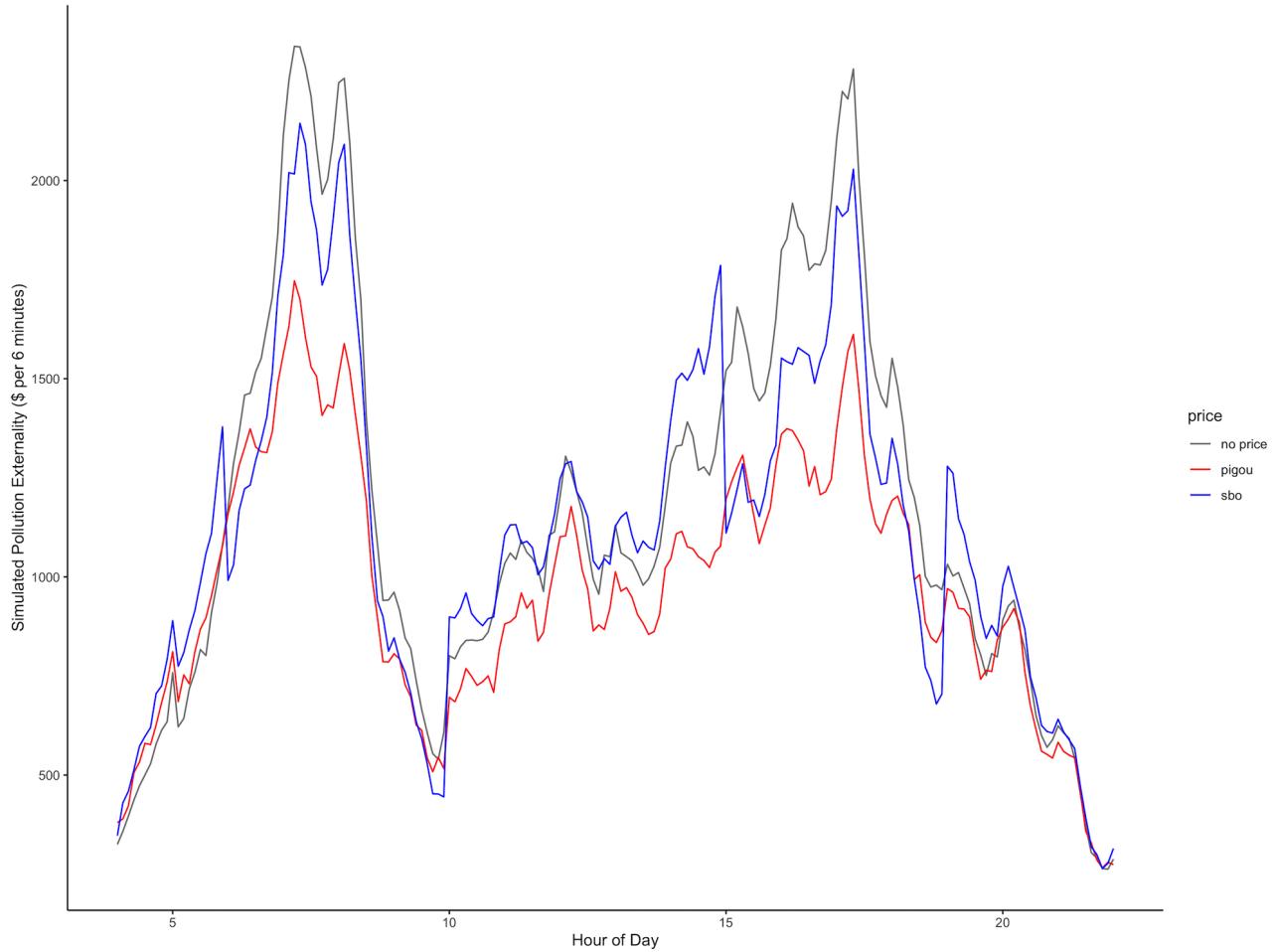


Figure 19: In this figure I plot the total pollution externalities under simulations using the mixed logit model estimated in Table 2 of Section 7 together with the NHTS trip dataset described in Section 8. In each scenario, I predict 600,000 choices — roughly daily total of vehicle trips that pass through 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 first-best policy where every trip a driver could choose (including non-cordon trips) would be priced according to its marginal pollution and congestion externalities. The red line plots sum of pollution externalities under the second-best optimal peak-hour cordon price from Figure 11. Note that all lines include pollution from trips that cross through the cordon, as well as “detour” trips that circumvent the cordon.

C. 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 (PM_{2.5}, or PM), nitrogen oxides (NO_x), nitrous oxide (N₂O), reactive organic compounds (ROC), ammonia (NH₃), carbon dioxide (CO₂), sulfur oxides (SO₂), and methane (CH₄). 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 E) 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 ₂ O	18,000
ROC	2,392

Table 11: This table display the social costs of emitting 1 ton of various pollutants in San Francisco. Estimates of local pollutants (PM_{2.5}, nitrogen oxides (NO_x), nitrous oxide (N₂O), reactive organic compounds (ROC), ammonia (NH₃), sulfur oxides (SO₂)) reflect annual averages from the EAISUR model ([Heo, Adams, and Gao, 2016](#)). Global pollutants (carbon dioxide (CO₂) and methane (CH₄)) 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 ₂ O	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, nitrogen oxides (NO_x), nitrous oxide (N₂O), reactive organic compounds (ROC), ammonia (NH₃), sulfur oxides (SO₂)) reflect annual averages from the EAISUR model ([Heo, Adams, and Gao, 2016](#)). Global pollutants (carbon dioxide (CO₂) and methane (CH₄)) 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 ₂ O	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, nitrogen oxides (NO_x), nitrous oxide (N₂O), reactive organic compounds (ROC), ammonia (NH₃), sulfur oxides (SO₂)) reflect annual averages from the EAISUR model ([Heo, Adams, and Gao, 2016](#)). Global pollutants (carbon dioxide (CO₂) and methane (CH₄)) are values used by the US EPA.

D. Bunching Estimator

Decomposition of the Bunching Estimator

Table 14 — BUNCHING ESTIMATOR FOR SCHEDULING COSTS (SHIFTING LATER, 10 A.M.)

Parameter	Estimate
Fraction Unresponsive (a)	0.87357
Excess Mass at Notch (B)	0.00136
Baseline Density at Notch, h_0	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 21). Rows 4-6 show estimates of scheduling costs. In Row 4, I calculate the naive average scheduling cost 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. In row 6, I also account for the difference in travel times for drivers who reschedule their trips to avoid peak-hour pricing.

Table 15 — BUNCHING ESTIMATOR FOR SCHEDULING COSTS (SHIFTING EARLIER, 5 A.M.)

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 21). Rows 4-6 show estimates of scheduling costs. In Row 4, I calculate the naive average scheduling cost 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. In row 6, I also account for the difference in travel times for drivers who reschedule their trips to avoid peak-hour pricing.

FIGURE 20 — TRAVEL TIMES IN THE VICINITY PRICE NOTCHES

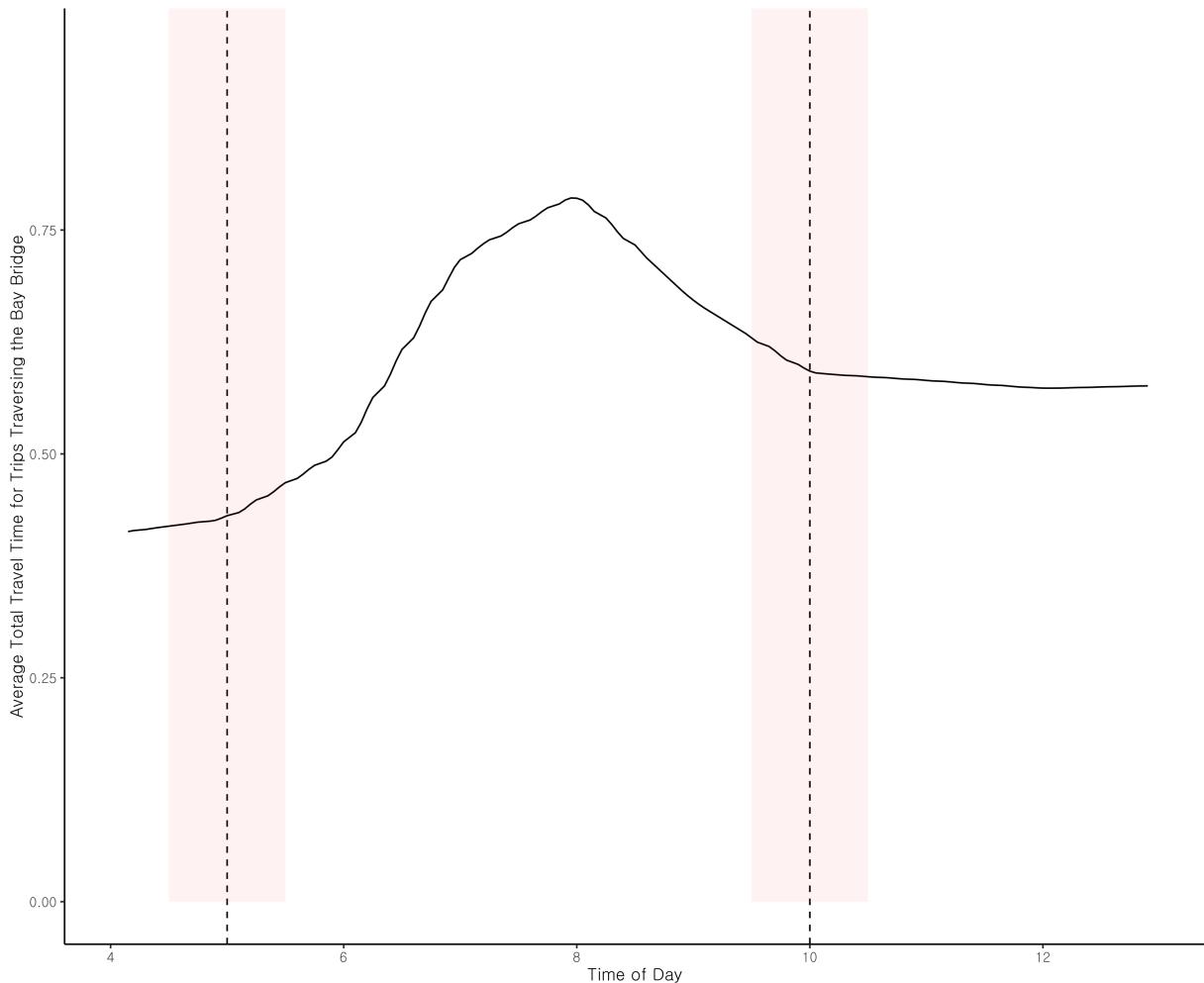


Figure 20: This figure plots average travel times for trips traversing the Bay Bridge during the morning hours. The average travel times in this figure were calculated by 1) identifying all drivers that primarily use the Bay Bridge and b) using TomTom Historic Traffic Stats to calculate travel times for each individual, for each hour of day as described in Section 5. The red shaded area represent the approximate range where individuals adjust in response to the imposition of peak-hour pricing, according to FasTrak toll data. 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, as opposed to changes in total travel time. As shown in Tables 14 and 15, estimates that account for differences in travel times in the bunching estimators are roughly 30% larger than estimates that ignore differences in travel times.

E. Imputing 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, $\{4.0, 4.2, \dots, 22\}$. 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). This survey constitutes a representative sample of Bay Area commuters, and contains 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 21.

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.

[This space is left intentionally blank]

FIGURE 21 — TOMTOM TRAFFIC SEGMENTS



Figure 21: This figure plots the coverage of the historic travel time data purchased from TomTom (in red) together with the 29 most populous cities in the Bay Area. These road segments were selected using Google Maps suggested driving points between the origin and destination cities. These traffic data report the average weekday travel times for passenger vehicles traveling along each segment of road, by hour of day, for the year prior (July 1, 2009 - July 1, 2010) and the year following (July 1, 2010 - July 1, 2011) the 2010 adjustment to Bay-Area bridge tolls.

F. Equilibrium Considerations

The second-best cordon price results presented in Section 8 reflect social damages calculated using traffic conditions in untaxed equilibrium. Consistent with the literature on externality taxation, the second-best tax formula in Section 2 phrases optimal taxes as a function of externalities at the optimum. As shown in figures 4 and 25, the marginal damages associated with driving are not constant in traffic density/speed, meaning that in general, damages at the taxed equilibrium will be different (lower) than those observed in the untaxed equilibrium. Whether the difference between marginal damages calculated at versus away from the optimum is a first-order concern depends on the slope of the marginal damages function and the responsiveness of drivers to taxation.

In this appendix, I simulate changes in traffic density under taxation to estimate a lower bound for the second-best optimal cordon prices in San Francisco. Specifically, I iteratively calculate traffic density, driver choices, and taxes until I reach a fixed point where driver's decisions under a given tax vector, τ^* , imply traffic densities (and associated externalities) such that applying equation 10 yields τ^* .

This algorithm is as follows:

1. Calculate cordon taxes (τ_n) using equation 10 and the externalities under current conditions (ϕ_0), as described in Section 8.
2. Use the NHTS dataset to simulate 600,000¹⁵ driver choices under τ_n .
3. Re-scale the hourly sensor-level road densities by comparing the simulated number of trips that would pass over a given sensor in a given hour under the status quo to the number of trips that would pass over a given sensor in a given hour under the simulation from step (2).
4. Re-estimate the social damages (ϕ_n) associated with each trip according to these updated hourly traffic densities.
5. Repeat steps 1-4, replacing externalities under current conditions (ϕ_0) in step 1 with ϕ_n , until the optimal cordon price in subsequent iterations meet some arbitrary element-wise threshold for convergence: $|\tau_n^h - \tau_{n+1}^h| < \epsilon$, where τ^h is element h of the tax vector calculated in step n .

Figure 22 plots the results of applying this algorithm to cordon pricing in San Francisco using a convergence threshold of \$0.01. The initial points (iteration one) are the taxes calculated with trip-level damages that reflect current traffic conditions, and are therefore equivalent to the results shown in Section 8 (see row 1 of Table 6). After 9 iterations of recalculating traffic density and taxes, the morning and evening converge to \$1.59 and \$1.80, respectively.

The fixed point from this exercise constitutes a lower bound because it ignores “induced demand,” or “rebound:” marginal drivers who would have chosen not to take a trip in the absence of road pricing, but choose to take the trip under road pricing due to lower travel times. For any step $n > 1$ in the above algorithm, induced demand would imply traffic densities higher than those estimated by the discrete choice model. Induced demand would therefore attenuate the difference in traffic conditions between taxed and untaxed equilibria (Duranton and Turner, 2011). Optimal taxes at the true equilibrium therefore lie between the results presented in Section 8 and the fixed point presented here.

¹⁵ According to the San Francisco County Traffic Authority, roughly 600,000 vehicle trips cross San Francisco's proposed cordon daily.

FIGURE 22 — BOUNDING EQUILIBRIUM EFFECTS

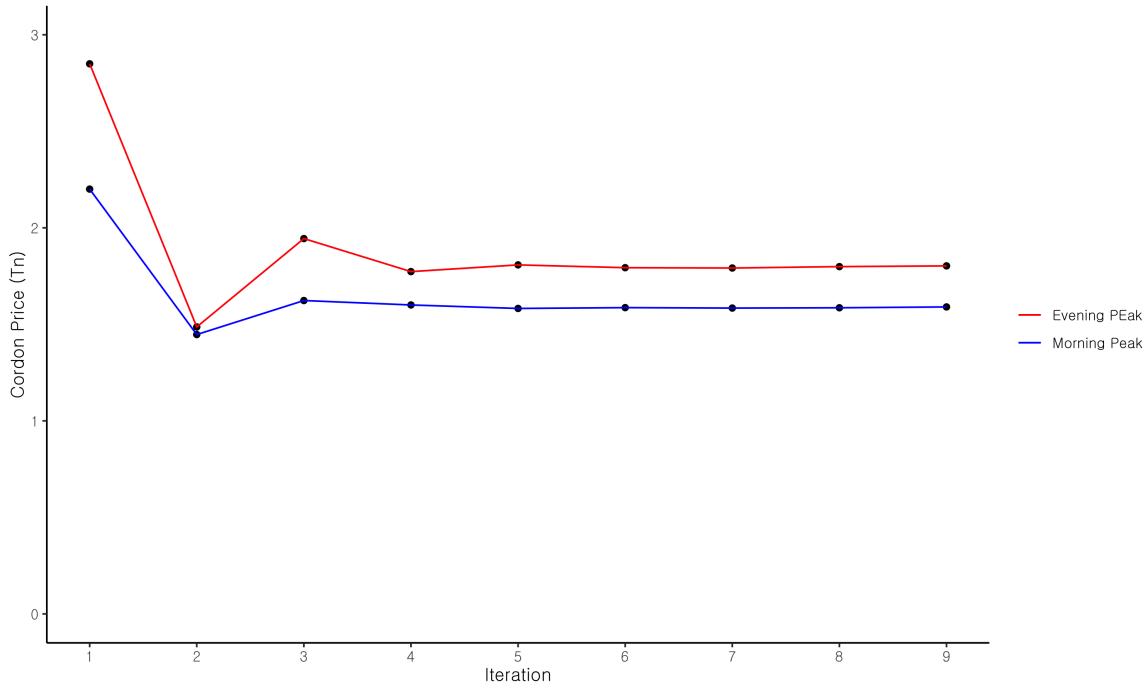


Figure 22: Per-mile driving externalities are larger under denser traffic conditions (Yang, Purevjav, and Li, 2020). As a result, trip-level traffic externalities calculated using untaxed traffic conditions may overestimate optimal taxes. This figure displays the results of the simulation exercise where I iteratively calculate traffic density, driver choices, and taxes until reaching a fixed point where driver's decisions under a given tax vector, τ^* , imply traffic densities (and associated externalities) such that applying equation 10 yields τ^* . This optimal tax contains two elements: morning (blue) and evening (red) peak hour prices. The fixed point in this exercise is a lower bound for the second-best peak-hour cordon prices in San Francisco because it ignores “rebound,” or “induced demand” — drivers adding other trips or shifting from other modes in response to the improved traffic conditions under taxation.

[This space is left intentionally blank]

G. Congestion Pricing and Accidents

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 risks of accidents in urban vs. rural areas. Empirical studies of the impact of congestion charges on accidents suggest that the value of accident reductions are several orders of 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.

H. 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.

H.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 driving 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:

$$\begin{aligned} \lambda &= 1 \\ u'(x) &= \phi(a) + p \\ \phi'(a)x &= c'(a) \end{aligned}$$

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:

$$\begin{aligned}\lambda &= 1 \\ u'(x) &= \phi(a) + c(a) + p \\ \phi'(a) &= c'(a)\end{aligned}$$

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.

The results in the body of this paper are not adjusted for existing environmental policies that impact the variable cost of driving, namely fuel content regulation. [Aufhammer and Kellogg \(2011\)](#) estimate that fuel content regulations in California cost roughly 12 cents (in 2020 dollars) per gallon. If an average trip crossing San Francisco's cordon boundary travels roughly 10 miles per hour and has a fuel efficiency of 20 miles per gallon, the second-best optimal prices in this paper adjusted for pre-existing fuel regulation would be roughly \$0.06 lower than the results shown in Section 8.

H.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.

I. External Validity

The appropriateness of using the driving 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, we use data from the 2017 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 23, for example, shows that similar fractions of drivers report flexible work schedules in each of these cities. Figure 24 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 25 through 27 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 23 — SCHEDULE FLEXIBILITY BY METRO AREA

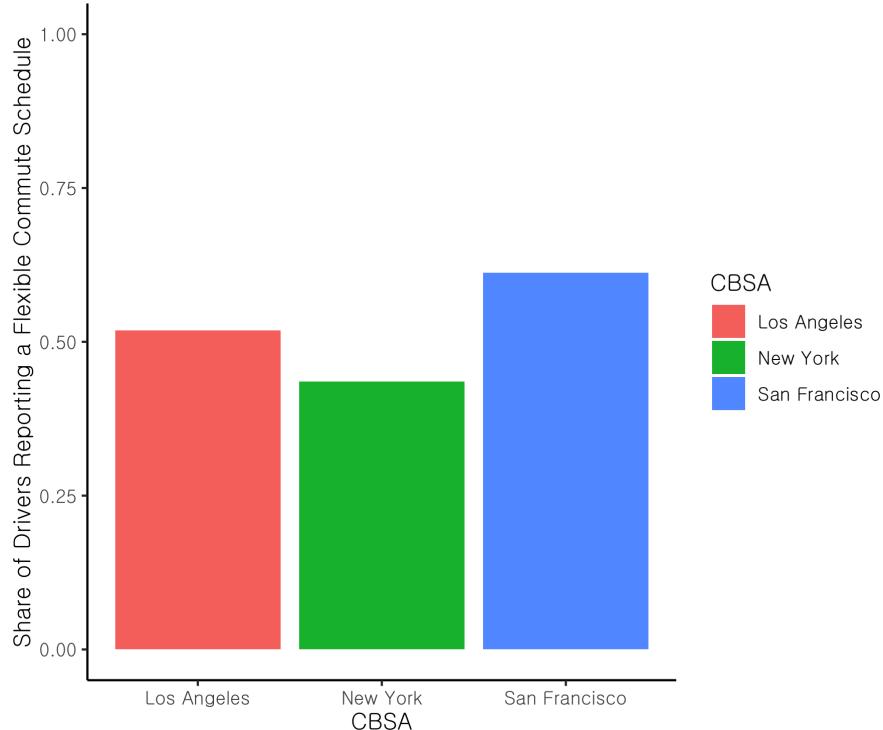


Figure 23: This figure plots the share of drivers who report having a flexible work schedule by metro area, according to the 2017 National Household Transportation Survey.

FIGURE 24 — SCHEDULE FLEXIBILITY BY TIME OF DAY

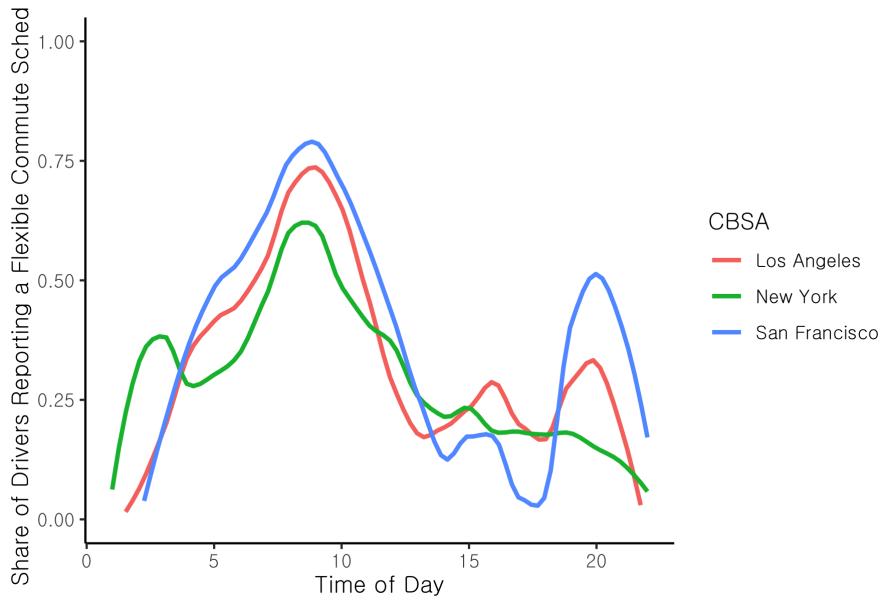


Figure 24: This figure plots the share of drivers who report having a flexible work schedule by time of day and metro area, according to the 2017 National Household Transportation Survey.

FIGURE 25 — EMISSIONS FACTORS VS. GAS PRICE RESPONSIVENESS

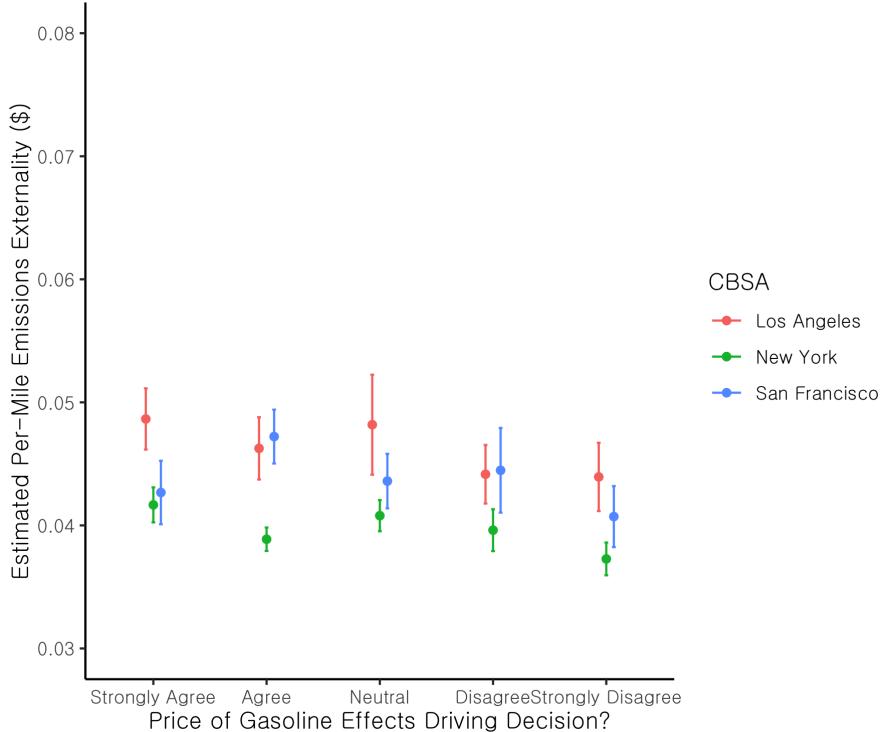


Figure 25: This figure plots estimates emissions factors of vehicles in the 2017 National Household Transportation Survey against vehicle owners' self-reported responsiveness of driving demand with respect to gasoline prices. Emissions factors reflect vehicle age and fuel type.

FIGURE 26 — VEHICLE AGE VS. GAS PRICE RESPONSIVENESS

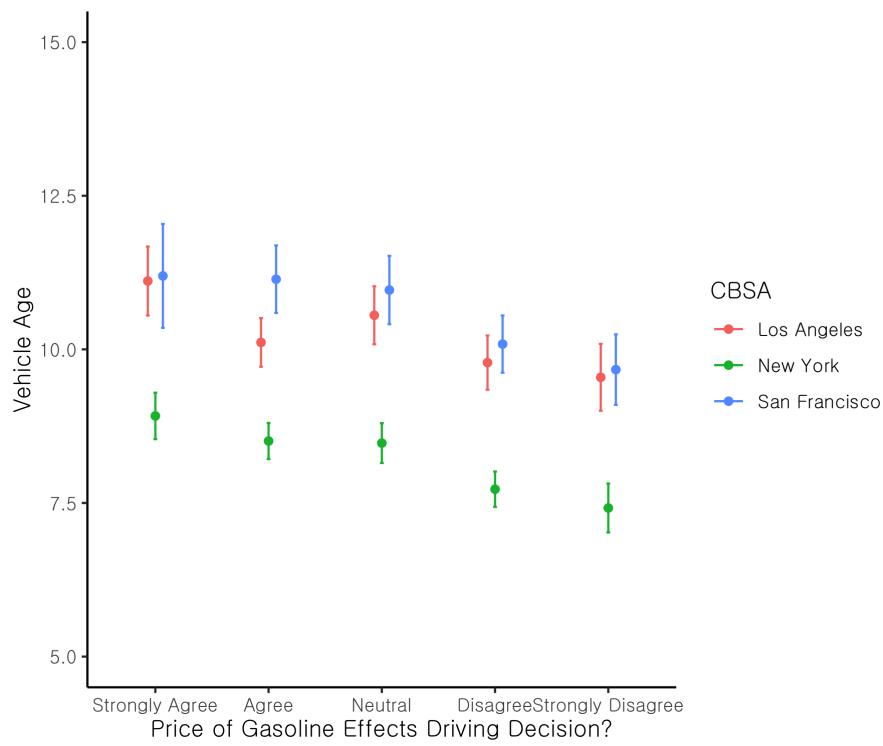


Figure 26: This figure plots vehicle age against vehicle owners' self-reported responsiveness of driving demand with respect to gasoline prices using data from the 2017 National Household Transportation Survey.

FIGURE 27 — CONGESTION VS. GAS PRICE RESPONSIVENESS

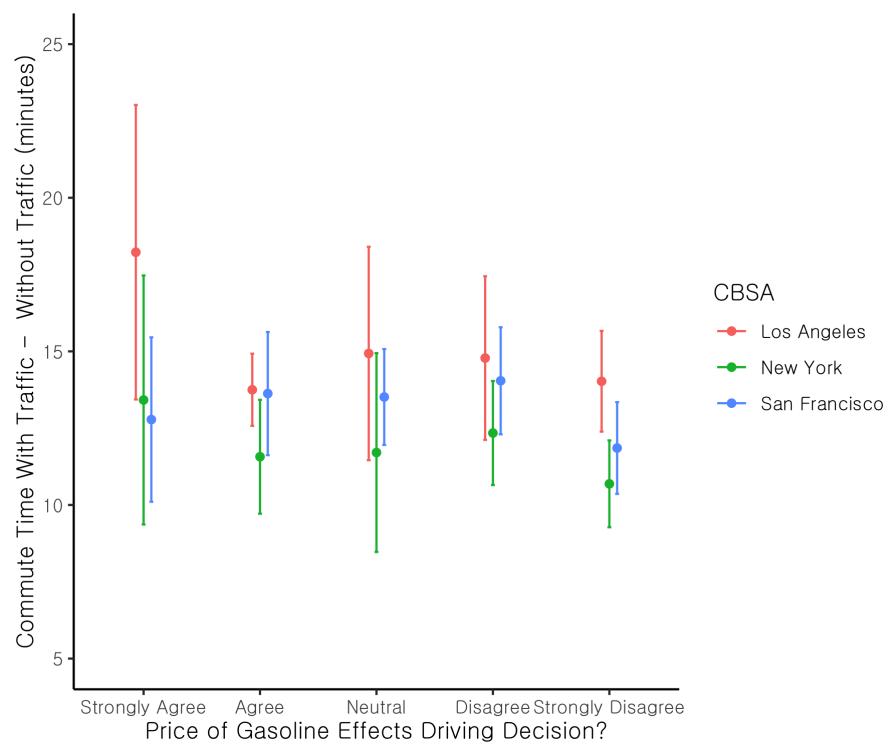


Figure 27: This figure plots self-reported gasoline price responsiveness against the amount of time a driver reports losing to traffic during their commute for drivers in the 2017 National Household Transportation Survey.

J. Low-Income Exemptions

While road pricing can increase economic efficiency, concerns about regressivity have prompted planners in many cities to consider road pricing schemes specifically designed to reduce the incidence on low-income road users. In San Francisco, for example, a majority of the congestion pricing proposals under consideration include some level of income-based exemption ([San Francisco County Traffic Authority, 2021](#)). Similarly, as of 2021, New York plans to refund congestion tolls for drivers who make under \$60,000 per year ([Regional Plan Association, 2021](#)).

Table 16 compares predicted reductions in pollution, congestion, and deadweight loss in San Francisco under (a) the second-best peak hour cordon prices estimated using equation 10, and (b) the same policy where drivers from low-income households (those with self-reported household income below \$75,000 in the NHTS) are exempt from cordon fees. This exercise suggests that the efficiency costs of these exemptions are modest: exempting low-income drivers from cordon pricing in San Francisco would generate reductions in pollution and congestion externalities that are 1-3 percentage points smaller than under an optimal no-exemption policy. These efficiency costs are substantially smaller than the efficiency costs that reflect other imperfections in cordon pricing. For example, I estimate that restricting pricing to peak hours generates reductions in congestion, pollution and deadweight loss that are 10 to 20 percentage points smaller than a pricing scheme where a policymaker can set a fixed hourly schedule of prices between 6 am and 10 pm (see Table 8).

The relatively small efficiency cost of exemptions reflects the low proportion of low-income drivers in trips that use the cordon. In the sample of 1,891 trips from the California NHTS with fastest routes that pass through California's cordon zone, just 9% are taken by drivers from houses with a total annual income of less than \$75,000.

Table 16 — SECOND-BEST CORDON PRICING WITH LOW-INCOME EXEMPTIONS

Outcome	Performance Relative to First-Best (%)	
	Second-Best Peak Hour	Low Income Exemption
Reduction in Total Externalities	30.576	28.173
Reduction in Congestion	31.367	28.782
Reduction in Pollution	23.779	22.939
Welfare Gain	32.304	32.601

Table 16: Column (1) of this table reproduces the results from Table 5, which compares outcomes under second-best optimal peak hour cordon pricing to outcomes under Pigouvian pricing. The second column in this table compares the first-best policy to cordon pricing scheme that is identical to the scheme column (1), except that households making less than \$75,000 per year are exempt from cordon fees. Income data reflect self-reported household income from the 2017 National Household Transportation Survey.