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

Matthew Tarduno[†]

September 9, 2024

Click here for the latest version

Abstract

Real-world congestion zones are imperfect because they charge heterogeneous road users uniform prices, and invite externality spillovers in space and time. I show that given these imperfections, calculating optimal prices requires (i) individual-level externalities, (ii) individual elasticities, and (iii) cross-price elasticities between priced and unpriced trips. Using bridge toll microdata and a natural experiment where peak-hour pricing was imposed on one of the San Francisco Bay Area's 4 trans-bay bridges, I estimate a discrete choice model of driving demand that yields these parameters. I then use this model to estimate optimal prices for proposed congestion zones in three U.S. cities. I find that leakage pushes second-best prices below trip-level externalities, and that optimal peak pricing recovers just 10-41% of the welfare gains of a first-best policy.

[†]University of Illinois at Chicago; tarduno@uic.edu. I thank James Sallee, Michael Anderson, and Reed Walker for their support throughout this project. I also thank Lucas Davis, Meredith Fowlie, Marco Gonzalez-Navarro, Alan Auerbach, Josh Apte, Max Auffhammer, Sofia Villas-Boas, Severin Borenstein, James Sears, Jenya Kahn-Lang, Marshall Blundell, Matthew Suandi, Elif Tasar, Eva Lyubich, Sara Johns, Matt Woerman, Marcus Casey and seminar participants at UC Berkeley, Williams College, NC State's Camp Resources, the European Meeting of the Urban Economics Association, 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.

1. Introduction

On June 5th, 2024, in a dramatic reversal, New York Governor Kathy Hochul made the decision to pause New York City's planned Congestion Pricing Program. Although the decision to halt congestion pricing was ultimately political, it reignited the policy debate over how New York's policy — if it is ever implemented — ought to be designed. Was the $15 toll favored by the MTA the right price? Or was Hochul right to suggest lowering the charge? Similarly, a number of policy attributes that appeared settled, from peak-hour toll designs to alternative exemption schemes, suddenly seemed in play again.

Surprisingly, existing research in urban and environmental economics offers limited prescriptions for New York policymakers. While economists have long advocated for charging road users to address the negative externalities associated with urban driving (Vickrey, 1963; Parry, 2002), 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; Selmoune et al., 2020). 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 relationship between these 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 theory, the two policy imperfections — leakage and heterogeneity — have an ambiguous effect on optimal cordon prices relative to prices in a first-best policy. Depending on the correlation between social damages and driver elasticities, heterogeneous externalities could imply optimal prices that are either above or below social damages (Diamond, 1973). The discrete spatial and temporal cutoffs in cordon pricing incentivize some drivers to shift trips in time and space to avoid tolls. Absent heterogeneity, this leakage would imply optimal prices that are unambiguously below social marginal damages (Green and Sheshinski, 1976). Whether second-best prices are above or below average Pigouvian prices therefore depends on the sign and strength of the correlation between price-responsiveness and individual externalities, as well as the size of the leakage effect.

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. In my main set of empirical results, 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. Additionally, I corroborate a subset of the parameter values from the discrete choice model using a bunching estimator: The temporal cutoff in the Bay Area's peak-hour tolls yields a notched budget constraint for drivers; the size of the bunching just outside of peak hours is proportional to the willingness of drivers to substitute trips in time.

Taking a discrete choice approach 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 rather than individually estimating substitution between each hour of day and possible route taken by a driver.

I use this model of driving demand to calculate second-best optimal prices for proposed cordon zones in San Francisco, Los Angeles, and New York. I find that leakage is the dominant consideration in calculating second-best cordon prices. The possibility of drivers substituting their trips in time or space leads to 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 $5.90 for the morning peak and $8.50 for the evening peak. These prices are below the average social damages generated by trips that use the cordon during those periods ($7.66 and $10.43, respectively). 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 41% of the total welfare gains relative to a policy where drivers are charged according to the marginal damages of each trip. Notably, these results are driven by the policy's ability to address congestion externalities, which tend to be 2 to 10 times larger on a per-trip basis than pollution externalities. Second-best peak-hour pricing performs even more poorly for proposed zones in New York (28% of the welfare gains of the first-best) and Los Angeles (10% of the welfare gains of the first-best).

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 $132 million annually in San Francisco to $514 million annually in New York. In each city, however, a cordon zone with second-best hourly prices would still leave a substantial portion (roughly 20 to 40%) of the possible welfare gains unrealized due to remaining issues of spatial leakage and imprecise pricing.

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 $5.90 - $8.50 in San Francisco to $11.30 - $16.30 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), 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 et al., 2012), policies designed without regard to agent re-optimization and heterogeneity may lead to poor or perverse policy outcomes (Davis, 2008; 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; Knittel and Sandler, 2018), which shows that second-best uniform prices are a weighted average of heterogeneous externalities, and the "leakage" model, where second-best optimal prices reflect marginal damages, less a term that captures leakage (substitution) to other unpriced goods that also generate externalities (Green and Sheshinski, 1976, see also Davis and Sallee, 2020; 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.

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 power plants (urban or rural; upwind or downwind of population centers), and policies implemented by states or utilities may allow for externality leakage if electricity is imported from other jurisdictions. Sin taxes (e.g., cigarette taxes) similarly have heterogeneous impacts on consumers, and taxing any single product may induce consumers to substitute towards related (and under-taxed) sin products (Fleissig, 2021).

Lastly, this paper presents a new bunching-based approach for estimating the willingness of commuters to reschedule their trips. Scheduling costs are key parameters in transportation economics (Vickrey, 1963; Arnott, De Palma, and Lindsey, 1990) and an important factor in determining the possible welfare gains from peak-hour congestion pricing (Kreindler, 2018). Adapting tools from the public finance literature on bunching, 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 other metro areas. Indeed, a close re-examination of existing empirical work shows this type of temporal bunching is quite common. As I detail in the results section, bunches around notches in congestion prices are visible in time-of-day plots of Milan (Gibson and Carnovale, 2015), Bergen (Isaksen and Johansen, 2021), Stockholm (Kristoffersson, 2013), and London (Transport

for London, 2008).

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, 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 policy. In Sections 2.1 and 2.2, I outline canonical models of optimal taxation under each of these separate imperfections (heterogeneity and leakage). Then, in Section 2.4, I 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 are charged the same price. 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.

Setup: consider n consumers that derive utility from their consumption of an externality-generating good, α_i, and disutility from other's consumption of this good: $U^i = U(\alpha_1, ..., \alpha_n) + \mu_i$. The second-best optimal uniform tax in this setting is:

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

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

This expression captures an important principle in second-best corrective taxation: If individual elasticities are positively (negatively) correlated with idiosyncratic externalities, the

second-best uniform tax on the externality-generating good will be larger (smaller) than the naive average of marginal damages. Intuitively, the role of corrective taxes is to move individuals to adjust their consumption of a product to the level where private marginal benefit equals the social marginal cost. If a given group is unresponsive to price, however, the second-best optimal tax described above will provide the correct incentive for the responsive group to consume at the level that balances private and social marginal costs.

2.2. Leakage

Just as legal or practical constraints prevent policymakers from perfectly targeting externalities, these constraints often also 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 (where one is taxable and the other 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 both the substitutability of the two goods, and the marginal damages of the untaxable good.

Consider a setting with two goods, x and y, and associated marginal external damages ϕ_x and ϕ_y, respectively. A representative consumer with an exogenous income derives utility from these two goods, and a quasilinear numeraire: $U = U(x, y) + z$. If a social planner is constrained to only tax x, then the optimal tax is:

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

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.

In the remainder of this section, I cover two extensions to the above models, the second of which characterizes optimal cordon prices.

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 untaxable. This intermediate setting provides intuition useful for understanding the model with heterogeneity presented in Section 2.4.

Setup: A representative consumer chooses quantities of M goods, $(h_1, ..., h_M)$ 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 $j \in \{1, ..., J\}$ where $J < M$. I assume goods $k \notin \{1, ..., J\}$ are un- or under-taxed.

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

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

This intermediate result is a generalization of the two-good case. Holding fixed all taxes other than τ_j, the optimal value for the final tax is its externality, ϕ_m, plus a term that captures consumer substitution to other goods, and the level of unpriced externality of those goods. Identifying the optimal tax vector requires simultaneously solving J equations in the form of Equation 3. To do so, one can rewrite Equation 3 to separate the tax and externality terms:

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

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

$$a_1^j \tau_1 + ... + a_l^j \tau_l + ... + a_J^j \tau_J = b_j \qquad \forall \quad j \in [1, J] \tag{4}$$

Where a_l^j and b_j are defined as: $\quad a_l^j = \frac{\frac{\partial h_l}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \quad$ and $\quad b_j = \phi_j + \sum_{m=1}^{M} \frac{\frac{\partial h_m}{\partial p_j}}{\frac{\partial h_j}{\partial p_j}} \phi_l$

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 two 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:

$$\underbrace{\begin{bmatrix} a_1^1 & ... & a_J^1 \\ & ... & \\ a_1^J & ... & a_J^J \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} \tau_1^* \\ ... \\ \tau_J^* \end{bmatrix}}_{\boldsymbol{\tau}} = \underbrace{\begin{bmatrix} b_1 \\ ... \\ b_J \end{bmatrix}}_{\boldsymbol{b}}$$

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

$$\boldsymbol{\tau} = \boldsymbol{A}^{-1} \boldsymbol{b} \tag{5}$$

Equation 5 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 ma-

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

2.4. Heterogeneity and Leakage

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

Although 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 externalities that vary by location (Muller and Mendelsohn, 2007), and local pollution policies may induce leakage if utilities or states import electricity across borders. Similarly, "sin" goods have 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 — heterogeneity is included in the model to reflect the implications of differences in externalities, rather than understand how externality taxation interacts with inequality aversion.

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

As in Section 2.3, a policymaker can choose tax levels for goods $j \in \{1, ..., J\}$ where $J < M$. I assume goods $m \notin \{1, ..., J\}$ are un- or under-taxed. I denote τ^j as the tax on good j. In Appendix A, I show that the optimal tax on τ_j as a function of the k other tax levels is:

$$\tau_j = \frac{\sum_{i=1}^{N} \sum_{g}^{N} (\frac{\partial U^i}{\partial h_g^1} \frac{\partial h_g^1}{\partial p_j} + ... + \frac{\partial U^i}{\partial h_g^M} \frac{\partial h_g^M}{\partial p_j})}{\sum_{i=1}^{N} \frac{\partial h_i^j}{\partial p_j}} + \frac{\sum_{k \neq j}^{J} \frac{\partial h_i^k}{\partial p_j} \tau_k}{\sum_{i=1}^{N} \frac{\partial h_i^j}{\partial p_j}} \tag{6}$$

This expression resembles the homogeneous case (Equation 3), but each of the externality terms is replaced by a Diamond-weighted term that accounts for individual-level heterogeneity in externalities. As in the previous case, the optimal tax vector solves a system of J equations:

$$\underbrace{\begin{bmatrix} a_1^1 & \dots & a_J^1 \\ & \dots & \\ a_1^J & \dots & a_J^J \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} \tau_1^* \\ \dots \\ \tau_J^* \end{bmatrix}}_{\boldsymbol{\tau}} = \underbrace{\begin{bmatrix} b_1 \\ \dots \\ b_J \end{bmatrix}}_{\boldsymbol{b}} \tag{7}$$

Now, a_l^j and b_j are defined as:

$$a_l^j = \frac{\sum_{i=1}^N \frac{\partial h_i^l}{p^j}}{\sum_{i=1}^N \frac{\partial h_i^j}{\partial p_j}} \qquad b_j = \underbrace{\frac{\sum_i^N \sum_{g \neq i}^N \frac{\partial U_i}{\partial h_g^j} \frac{\partial h_g^j}{\partial p_j}}{\sum_i^N \frac{\partial h_i^j}{\partial p_j}}}_{\text{Diamond-weighted externality of } j} + \underbrace{\sum_{l \neq j}^M \frac{\sum_i^N \sum_{g \neq i}^N \frac{\partial U_i}{\partial h_g^l} \frac{\partial h_g^l}{\partial p_j}}{\sum_i^N \frac{\partial h_i^j}{\partial p_j}}}_{\text{Diamond-weighted leakage shares}} \tag{8}$$

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 a discrete choice model 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.

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 $h \in H$ and routes $r \in R$ to satisfy their demand for travel. Included in this choice set is the outside (no trip) option. Each driver has an exogenous ideal arrival time, $\tilde{h^i}$. Drivers are atomistic and face travel times $T^i(h, r)$ 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 hour, respectively. Drivers also incur disutility α from each hour spent commuting. Lastly, each driver has idiosyncratic preferences for routes

and travel times, $\varepsilon^i_{h,r}$.[1] Driver i thus receives the following indirect utility from traveling via route r at time h:

$$u^i(h,r) = -\alpha T^i(h,r) - \gamma_e \underbrace{|h + T^i(h,r) - \tilde{h}^i|_-}_{\text{time early}} - \gamma_l \underbrace{|h + T^i(h,r) - \tilde{h}^i|_+}_{\text{time late}} - \beta p(h,r) + \varepsilon^i_{h,r} \quad (9)$$

To clarify the mapping between this discrete choice model and the optimal tax formula (Equation 7), 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*: $g \in H \times R$. Typical cordon zones have discrete spatial and temporal cutoffs.[2] 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. A purely reduced-form approach would require exogenous variation in congestion tolls for each possible route and hour of day to estimate how drivers substitute between available options. Alternatively, Equation 9 implies a matrix of own and cross-price elasticities between any choice set. These elasticities are a function of model primitives (α_e, γ_e, γ_l, β) and trip attributes ($T^i(h,r)$, p, *time late*, and *time early*). In Section 4 through 7, I use tolling microdata to recover estimates of each of these parameters using a multinomial 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.

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

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

[1] Whether idiosyncratic preferences are identically and independently distributed or correlated within nests of trips carries implications for substitution patterns implied by this model. In the empirical section, I present results under both of these assumptions, and adopt the nested approach as my preferred specification when using this model to estimate substitution patterns. See Section 6.1 for further discussion.

[2] The London Cordon Zone, for example, charges road users £15 between 7 am and 10 pm. The Milan Cordon Zone charges users €2 to €5 based on vehicle type between 7:30 am and 7:30 pm. San Francisco's proposed zone would only charge drivers during morning and evening peak hours.

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

Before 7/1/2010 *After 7/1/2010*

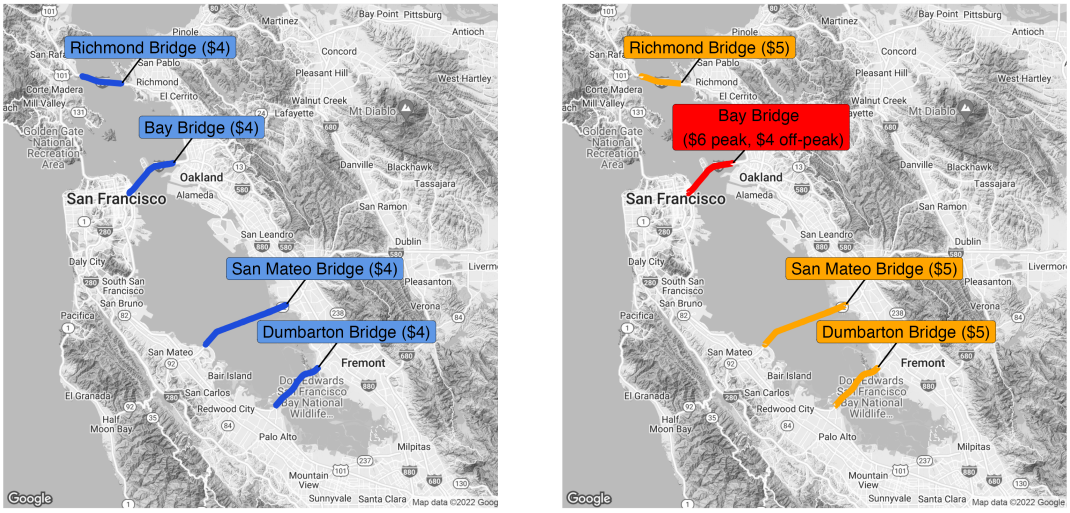

Figure 1: This map shows the four San Francisco Bay Area bridges used to estimate driver responses to toll prices in this paper. The *Richmond Bridge* connects Richmond and the eastern Bay Area to San Rafael and Marin County. The *Bay Bridge* connects Oakland to San Francisco. The *San Mateo Bridge* 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 A2).

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. Currently, passenger vehicles are charged between $3 and $7, depending on the time of day, the number of occupants, and whether or not the vehicle is electric/hybrid.

In this paper, I leverage several changes in the tolling structure that occurred on July 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 A2.

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. (*peak hours*) were charged $6. Tolls for all other hours (*off-peak*) remained at the pre-2010 price of $4.

Before 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.

Foreman (2016) uses reduced-form approaches to provide valuable estimates of the responses of Bay Area drivers to this change in bridge prices. They estimate 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, Foreman also finds that crossings on the San Mateo Bridge *increased* by 100 to 200 vehicles (around 5%) during peak hours, implying that some drivers switched from the Bay Bridge to 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 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 discrete choice 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 electronic trans-

actions 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.

In estimating the discrete choice model, I restrict the dataset on several dimensions. First, I include only devices with a valid Bay Area zip code. Second, for the purposes of estimation, I consider only weekday trips taken in a 2-week window before and after the 2010 change in toll prices (i.e., June 15^{th}, 2010 to July 15^{th}, 2010). Third, in estimation, I use only the morning commute hours (4 a.m. to 12 p.m.). Lastly, I drop devices with infrequent use (fewer than 2 weekday trips per week in the year prior to the 2010 price change), or users that take multiple westbound cross-bay trips per day during the 30-day study period. The resulting pool consists of 8,927 FasTrak devices and 78,226 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 reflect 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 of 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 the 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, $\tilde{h^i}$ in Equation 9, are not directly observed, and therefore must be inferred from each driver's activity. For each driver, I assign $\tilde{h^i}$ 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 to just outside of the peak pricing window.

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 a given trip and the idiosyncratic externalities associated with that trip (see Section 2). I therefore estimate the congestion and pollution externalities associated with each FasTrak trip.

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 exter-

nalities 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 cities—results in relatively few fatal traffic accidents. I provide further discussion of 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 A3.

Using a 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 A9. 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/mile)[3] 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/mile) along the route, times the length of the trip.

As noted above, because one trip termini is missing from the FasTrak data, I impute the congestion externalities for the missing segment of the trip (between the bridge and 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.

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

[3]The estimates from Yang, Purevjav, and Li (2020) are in yuan/vehicle/km. I convert these values to dollars by (a) converting currencies, and (b) replacing 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 et al. (2020).

(NO_x), and reactive organic compounds (ROCs), which negatively impact the health of nearby residents (Anderson, 2020; Currie and Walker, 2011). Vehicle emissions factors—the amount of a particular pollutant that a vehicle emits while traveling a mile—depend on a number of variables, including type of fuel consumed, fuel economy, vehicle age, and vehicle speed.

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. For global pollutants, I use 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.

Together, the data outlined in this section allow me to recreate the choices and choice sets facing a sample of FasTrak users, as well as the social costs associated with these choices.

6. Empirical Strategy

I use two strategies to recover the primitives that determine driving behavior. First, I use the variation in toll prices in 2010 together with the FasTrak microdata to estimate 9 via ordinary multinomial logit regression and nested logit regression. As a check for these results, I apply a bunching estimator to the Bay Bridge's notched tolling, producing a second estimate of scheduling costs.

6.1. Multinomial Logit and Nested Logit Regressions

As described in Section 5, the FasTrak microdata and the TomTom historic traffic data allow me to reconstruct the attributes of alternatives in the choice set for each driver. Namely, the different bridges, travel times, and prices facing trans-bay commuters. I then use this reconstructed choice set to estimate the discrete choice model of driving demand outlined in Section 3. Specifically, I estimate several variations of the following multinomial logit regression:

$$\mathbb{1}^i_{r,h,d} = \alpha \cdot time^i_{r,h,d} + \gamma_e \cdot early^i_{r,h,d} + \gamma_l \cdot late^i_{r,h,d} + \beta \cdot price_{r,h,d} + X_{r,h,d} + \epsilon^i_{r,h,d} \tag{10}$$

The dependent variable is an indicator for whether individual i chose to travel on route r at time of day h on date d. $time^i_{r,h,d}$ is the travel time associated with a route and time of day for individual i, $early^i_{r,h,d}$ is the time early at arrival, and $late^i_{r,h,d}$ is the time late. The $price_{r,h,d}$ variable does not vary by individual, but varies by bridge, time, and date. $X_{r,h,d}$ is a set of fixed effects for peak hours on each bridge, which I discuss below. $\epsilon^i_{r,h,d}$ is an extreme value error term. It is either independent and identically distributed (multinomial logit) or correlated within groups of alternatives (nested logit).

My preferred specification includes two adjustments to equation 10 made to meet the demands of the optimal tax formula: First, I interact the *price* variable with the externality quartile. Doing so produces estimates of price-responsiveness that are different for drivers who generate different amounts of externalities on their trips. Second, to allow for more flexible substitution patterns than those assumed by an ordinary logit model, I estimate a nested logit model with two nests: one degenerate nest that includes the outside "no trip" option, and a second nest that includes all possible bridges and times of day.[4] I further discuss the IIA assumption in this context in the following subsection.

Identification of the parameters of interest (β, γ_e, and γ_l) requires that the idiosyncratic error term is uncorrelated with choice attributes. The design of the peak-hour pricing on the bay bridge serves as a potential threat to identification: if trips during peak hours are popular for reasons other than the distribution of ideal arrival times (e.g. if there is a benefit to agents to having a similar schedule to other members of their household), then the peak-hour pricing in the post period could create a mechanical correlation between bay-bridge prices and idiosyncratic errors. To address this concern, I include peak-hour fixed effects for each bridge, which would subsume any temporal correlation between idiosyncratic errors and prices. These fixed effects do not absorb all of the variation in trip price; after the inclusion of these fixed effects, the price coefficient reflects differences in trip behavior before versus after the change in tolls, as well as extensive margin responses.

6.2. Own and Cross-Price Derivatives

The empirical strategies described above yield estimates of each of the primitives in Equation 9. Recall from Section 2 that the optimal cordon price formula requires information about how individuals substitute between available trip options. Given a set of routes and trip attributes, these parameter estimates imply a matrix of own- and cross-price derivatives. I can therefore use the parameter estimates from Equation 10 to predict substitution behavior in counterfactual settings. For goods x_j and x_k, the own and cross-price derivatives implied by a multinomial logit regression within an externality quantile are:[5]

Ordinary multinomial logit:

$$\partial s_j / \partial p_k = \begin{cases} \beta s_j (1 - s_j), & \text{if } j = k \\ \beta s_j s_k, & \text{otherwise} \end{cases} \tag{11}$$

[4]The two-nest logit model with a degenerate nest is a common approach in settings when there is a concern that cross-price elasticities between goods may differ from the cross-price elasticity of goods and the outside option. See, for example, Bertoli, Moraga, and Ortega (2013), Sheu (2014), or Kovach and Tserenjigmid (2022).

[5]See Train (2009) and Conlon and Mortimer (2021) for ordinary and nested logit derivatives, respectively.

Nested logit logit:

$$\partial s_j/\partial p_k = \begin{cases} \beta s_k(\frac{-1}{1-\sigma} + \frac{\sigma}{1-\sigma}s_{j|G} + s_j), & \text{if } j = k \\ \beta s_k(\frac{\sigma}{1-\sigma}s_{j|G} + s_j), & \text{if } j,k \text{ in the same nest, G} \\ \beta s_j s_k, & \text{otherwise} \end{cases} \tag{12}$$

Independence of irrelevant alternatives: Ordinary multinational logit regressions are correctly specified when the independence of irrelevant alternatives (IIA) assumption holds. In my preferred specification, I use a nested logit model with a degenerate nest for the outside option to partially relax the IIA assumption. I also present results with three nests: "no trip," "Bay Bridge," and "other bridges."

It is also worth discussing why the IIA assumption may not be as pernicious in this setting as in other instances where discrete choice models are used to study substitution. Broadly, this is because the optimal tax formula relies on own and cross-price *derivatives*, not *elasticities*. Under IIA, cross-price elasticities are equal across goods, but derivatives are not. Holding fixed the good experiencing a price change, the cross-price elasticity between that good and a candidate substitute is proportional to the choice probability of the substitute.

Said differently, nesting structure may be very important when many goods have very similar choice probabilities. In these cases, an ordinary logit may incorrectly predict very similar cross-price derivatives because shares are similar when in reality substitution depends more subtly on attributes. In the setting of peak-hour transportation with a popular outside option, however, the large differences in choice shares mean that in practice, the predicted cross-price derivatives are similar for nested and ordinary logit. Specifically, most of the mass in cross-price derivatives is focused on the outside good and prime-hour trips, regardless of the specification.

6.3. 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 not only acts as a check for the results from the logit regressions, but it is also a novel approach to estimating scheduling costs.

Bunching estimators are used to infer elasticities or other structural parameters from the empirical density of choice variables around kinks or notches in a budget set (Chetty et al.,

2011; Saez, 2010; Kleven and Waseem, 2013).[6]

Generally, 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 13 shows this indifference condition in terms of parameters from the discrete choice model, Equation 9. For simplicity, I examine the case of a driver who faces a decision of whether or not to shift their trip earlier:

$$\underbrace{\beta \Delta p}_{\text{Benefit from shifting}} = \underbrace{\gamma_e \Delta h}_{\text{Cost of shifting}} \tag{13}$$

Following the notation from Equation 9, β is the marginal utility of a dollar, 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:

$$\gamma_e = \frac{\beta \Delta p}{\Delta h} \tag{14}$$

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

$$\gamma_e = \frac{\beta \Delta p + \alpha \Delta T}{\Delta h} \tag{15}$$

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 2.

[6] Blomquist et al. (2021) note that under unrestricted preference heterogeneity, the size of bunching at kink/notch points does not identify an elasticity of taxable income. Although I do not explicitly model preference heterogeneity, there are two characteristics of my empirical setting that limit potential biases. First, using pre-period data, I directly observe the counterfactual density of Bay Bridge trips. Second, there is a natural limit to how far drivers are willing to shift their trips: The peak-hour periods are 4-5 hours long. This provides an upper bound for how far drivers shift to "bunch."

FIGURE 2 — THE RELATIONSHIP BETWEEN SCHEDULING COSTS AND BUNCHING

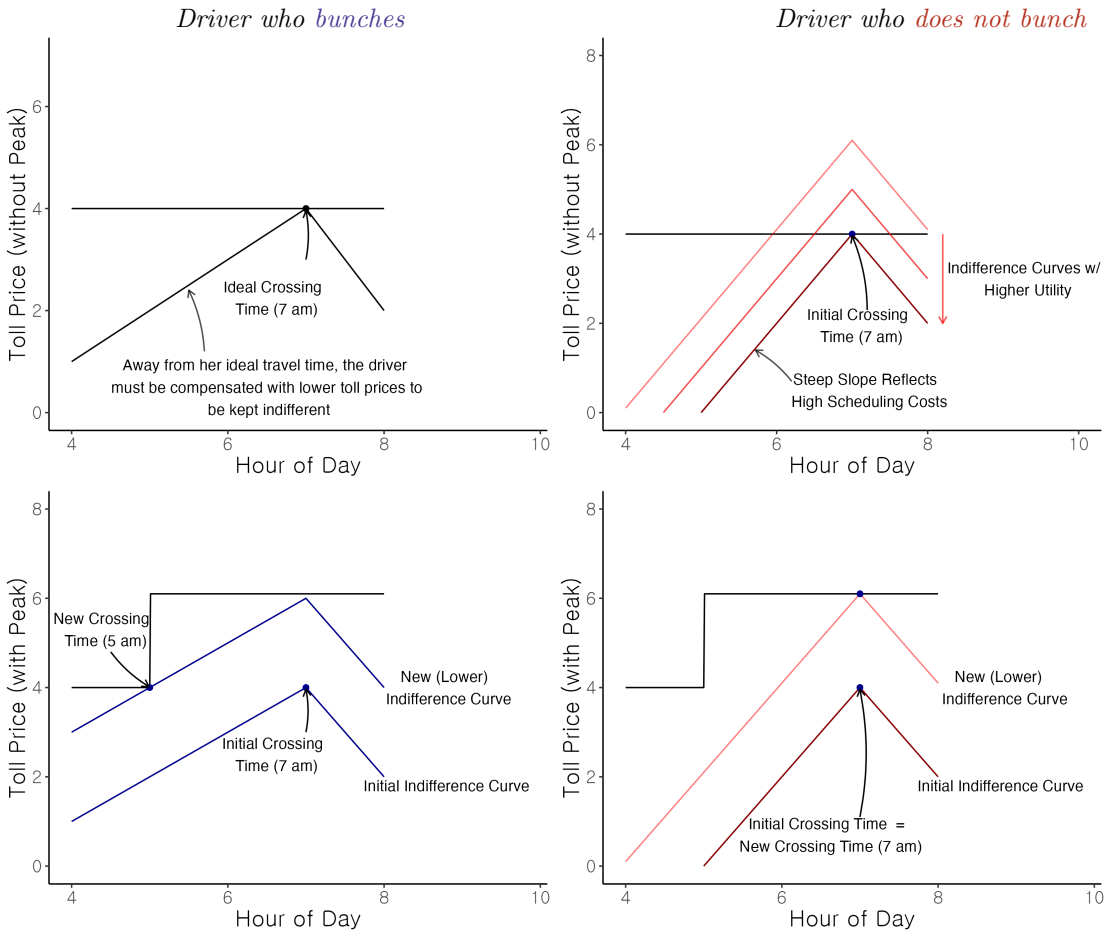

Figure 2: This figure illustrates the relationship between scheduling costs and bunching behavior in peak-hour toll schemes, as predicted by the discrete choice model outlined in Section 3. For expositional ease, this figure plots the case where travel times are constant throughout the day. The triangular 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 the 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.

Equations 14 and 15 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 A1) 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 (as in Finkelstein 2009) or the inability to perfectly time bridge crossings due to traffic shocks en route.

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:

$$B = \int_{\gamma_e} \int_{h*}^{h^*+\Delta h} (1-a) f_0(h) dh \simeq (1-a) f_0(h^*) E[\Delta h] \tag{16}$$

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 16 for Δh and plugging into Equation 15 yields the bunching estimator:

$$\gamma_e = \frac{\beta \Delta p}{B/((1-a) f_0(h^*))} \tag{17}$$

Relaxing the assumption that travel times are relatively flat around the notch point is straightforward but necessitates the value of travel time:[7]

$$\gamma_e = \frac{\beta \Delta p + \alpha \Delta T}{B/((1-a) f_0(h^*))} \tag{18}$$

7. Results

7.1. Reduced-form results

Before discussing the results from the discrete choice model that I use to calculate second-best cordon prices, this subsection briefly presents reduced-form evidence of drivers' responsiveness to the change in tolls in the four major trans-bay bridges in the Bay Area. The reduced-form regressions I estimate should not be taken as the main empirical results in the paper. They serve as a series of simple empirical results that motivate the approach that follows, and are broadly similar to previous studies of this policy (see Foreman, 2016).

[7] 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 17) in Appendix D.

In Figure 3, I plot estimates of changes in total traffic on each bridge, by hour of day. Each point and confidence interval in each panel of this figure reflects the results of applying equation 19 to aggregate traffic volume data from the corresponding bridge-hour. Tables 1 displays the results of regression discontinuity models where traffic data are aggregated up to the day level; 2 performs the same exercise, but restricts the aggregation to peak hours (6-10 am, and 3-7 pm).

$$count_{r,h,d} = \alpha_0 + \beta_1 \mathbb{1}(d > c) + \beta_2 \mathbb{1}(d > c)(d - c) + \beta_3 \mathbb{1}(d > c)(d - c) + \Gamma X_d + \epsilon_{r,h,d} \quad (19)$$

Two high-level conclusions emerge from these reduced-form exercises. First, point estimates suggest that traffic volumes decreased in response to increased toll prices. The only large positive point estimates are for off-peak hours on the Bay Bridge (prices for trips taken on the Bay Bridge during these hours did not change before versus after July 1, 2010). Second, the closest substitutes for the Bay Bridge are the San Mateo Bridge and Dumbarton Bridge. Although point estimates for peak-hour traffic changes are negative for each of these bridges, these estimates are not statistically significant. The point estimates are also smaller in relative magnitude than the decrease observed on the San Rafael Bridge, which experienced the same 1-dollar increase in prices.

While these results are consistent with a story where commuters substitute between routes and times of day in addition to extensive margin responses to tolls, we cannot distinguish between these responses from the aggregate toll crossing data alone. This shortcoming motivates the discrete choice approach that follows.

Table 1 — Reduced-form estimates of changes in hourly bridge traffic

	Bay Bridge	San Mateo Bridge	Dumbarton Bridge	San Rafael Bridge
Post	-5349**	-2643	-2976*	-2606**
	(1984)	(1852)	(1557)	(993)
Linear Trend (Pre)	17	-29	-61***	-45**
	(34)	(27)	(14)	(18)
Linear Trend (Post)	350***	298**	259**	182***
	(114)	(113)	(96)	(65)
Mean	107287	34298	25872	29261
Day of Week FE	✓	✓	✓	✓

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1: This Table shows estimates of change in westbound traffic on each of the San Francisco Bay Area's four trans-bay bridges following the July 2010 change in tolls. Each coefficient reflects applying equation 19, a regression discontinuity specification, to hourly data from the corresponding bridge. The coefficient in each column is the estimated change in total hourly trips. The mean number of hourly trips is included in the final row of the table for reference. Heteroskedasticity-robust standard errors are reported in parentheses.

Figure 3 — Impact of the 2010 Toll Changes on Bay Area Bridge Traffic

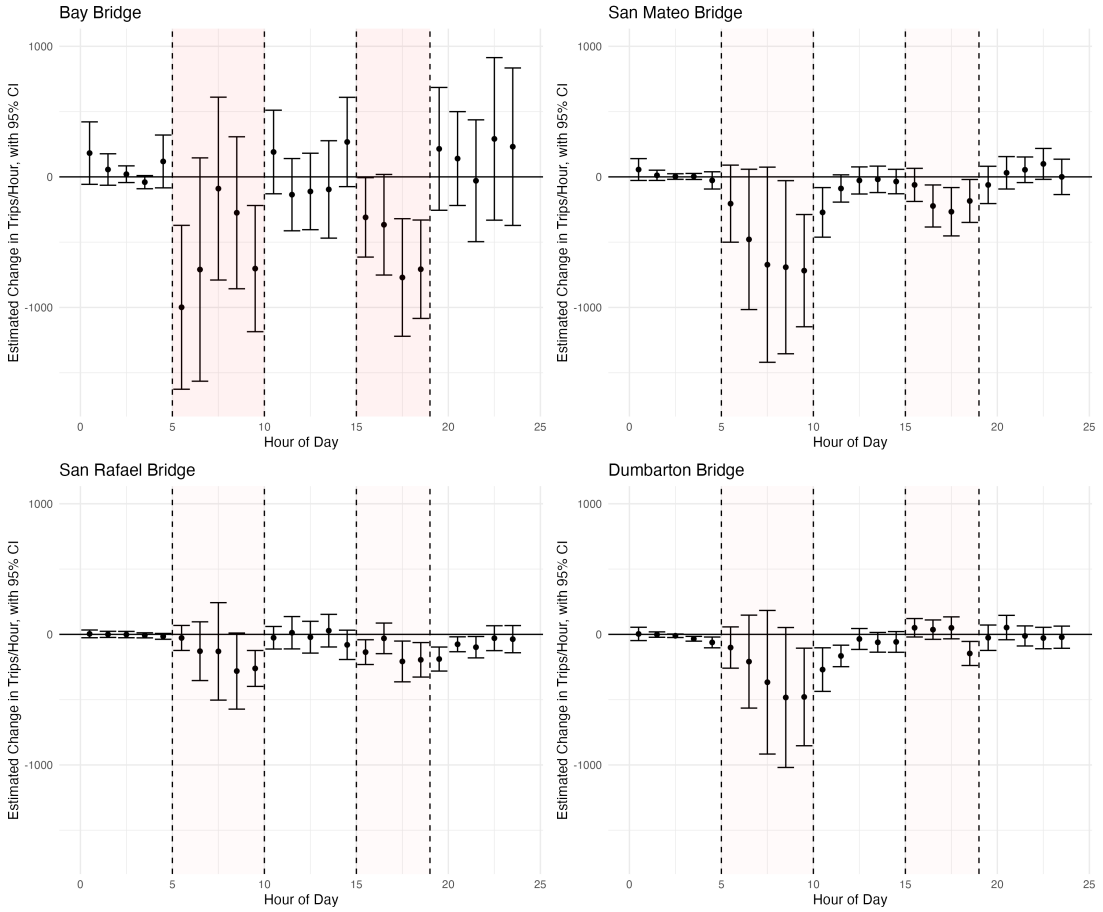

Figure 3: This figure displays hourly estimated changes in traffic volume on the four major trans-bay bridges after the 2010 change in toll prices. Each point represents the estimated coefficient for a given bridge an hour of the day, as per equation 19. The red shaded region corresponds to the peak hour pricing on the Bay Bridge; the other three bridges did not have peak hour pricing either before or after the toll change. The data used to arrive at these estimates are hour-by-date aggregate bridge crossing data courtesy of FasTrak.

Table 2 — Reduced-form estimates of changes in traffic, peak hours only

	Bay Bridge	San Mateo Bridge	Dumbarton Bridge	San Rafael Bridge
Post	-4993**	-2285	-2078	-1988*
	(2377)	(1827)	(1475)	(1073)
Linear Trend (Pre)	-34	-8	-43***	-41**
	(40)	(24)	(15)	(17)
Linear Trend (Post)	344**	215*	209**	164**
	(136)	(112)	(90)	(67)
Mean	47148	19218	15165	15560
Day of Week FE	✓	✓	✓	✓

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 2: This Table repeats the regressions from Table 1, but restricts the sample to only peak hours. Recall that peak hours — 6 a.m. to 10 p.m. and 3 p.m. to 7 p.m. — are the hours where prices on the Bay Bridge increased from $4 to $6. The variation in bridge tolls is explained in detail in Section 4. Each coefficient reflects applying equation 19, a regression discontinuity specification, to hourly data from the corresponding bridge. The coefficient in each column is the estimated change in total hourly trips. The mean number of hourly trips during the peak period is included in the final row of the table for reference. Heteroskedasticity-robust standard errors are reported in parentheses.

7.2. Discrete Choice Model Results

Table 3 presents the parameter estimates for the discrete choice model of driving demand (Equation 9) estimated via multinomial logit and nested logit regressions. The first column shows the model estimated with a uniform price-responsiveness. The coefficient on *travel time*, *time early*, and *time late* can be converted into units of dollars per hour by dividing the point estimate by the coefficient on *price*. Column 1 therefore suggests that drivers are indifferent between saving roughly $23.06 and saving an hour of travel time; they are indifferent between saving roughly $15.94 arriving an hour early, and they are indifferent between saving roughly $11.75 arriving an hour late.

In Column 2 of Table 3, I allow price responsiveness to vary with road users' idiosyncratic externalities. To do so, I break FasTrak devices into quantiles based on the average estimated externality (both pollution and congestion) of each device's choices in the month prior to the study period. I find little evidence that price elasticity varies systematically with idiosyncratic externalities in this sample.

Column 3 of Table 3 is my preferred specification. It is a nested logit model with two nests — a degenerate nest that includes the outside option (no trip) and a second nest that includes all bridges and routes. The results are similar to those from the multinomial logit: The implied value of travel time estimates are roughly $20 to $30, depending on the externality quantile, and

the per-hour scheduling costs are between half and two-thirds of the per-hour value of travel time. The final specification, Column 4, adds a third nest for the region's busiest bridge, the Bay Bridge. The parameter estimates are similar, with somewhat larger price coefficients in relative terms. As a result, the implied value of time ($14.95 to $16.81) and scheduling elasticities ($10.85 to $6.67) are lower.

Table 3 — Discrete Choice Model Estimates

	Multinomial (Pooled)	Multinomial	Nested (2 Nests)†	Nested (3 Nests)
Travel Time	-3.69^{***}	-3.69^{***}	-0.65^{***}	-2.69^{***}
	(0.01)	(0.01)	(0.03)	(0.02)
Time Early	-2.55^{***}	-2.55^{***}	-0.38^{***}	-1.74^{***}
	(0.01)	(0.01)	(0.02)	(0.03)
Time Tate	-1.88^{***}	-1.88^{***}	-0.27^{***}	-1.20^{***}
	(0.01)	(0.01)	(0.01)	(0.02)
Price	-0.16^{***}			
	(0.02)			
*Price * Quantile 1*		-0.16^{***}	-0.03^{***}	-0.18^{***}
		(0.03)	(0.00)	(0.01)
*Price * Quantile 2*		-0.18^{***}	-0.02^{***}	-0.17^{***}
		(0.03)	(0.00)	(0.01)
*Price * Quantile 3*		-0.12^{***}	-0.02^{***}	-0.16^{***}
		(0.03)	(0.00)	(0.01)
*Price * Quantile 4*		-0.19^{***}	-0.03^{***}	-0.17^{***}
		(0.03)	(0.00)	(0.01)
Nest: Any Trip			0.15^{***}	
			(0.01)	
Nest: Other Bridge				0.70^{***}
				(0.01)
Nest: Bay bridge				0.83^{***}
				(0.01)
Peak x Bridge FE	✓	✓	✓	✓

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 3: This Table contains parameter estimates for variations Equation 9, a discrete choice model of travel behavior described in Section 6. 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 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. The choice data used to estimate these equations are individual-level FasTrak records from the San Francisco Bay Area; the travel time data are derived from historical TomTom traffic data. † is an indicator for my preferred specification, and the specification I rely on when calculating optimal cordon prices.

7.3. 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 9) that range from \$6 to \$15 per hour.

Figure 4 plots 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 5 a.m. bunch in the post-period density of trips 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.

Figure A6 plots the frequency of vehicle trips 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 to arrive late for both morning and evening trips.

Using Equation 18, 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 4.

Table 4 — Estimating Scheduling Costs via Bunching

Parameter	**Estimate (\$\hour)**
Time Early (5 a.m. notch)	6.195
	(0.419)
Time Early (3 p.m. notch)	9.744
	(1.545)
Time Late (10 a.m. notch)	15.498
	(2.593)
Time Late (7 p.m. notch)	13.759
	(1.306)

Table 4: 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 9). I recover these estimates using Equation 18, 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 18 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 16) using administrative tolling data from the Bay Area Toll Authority. Bootstrapped standard errors are in parentheses.

Figure 4 — Bunching in Response to Peak-Hour Pricing

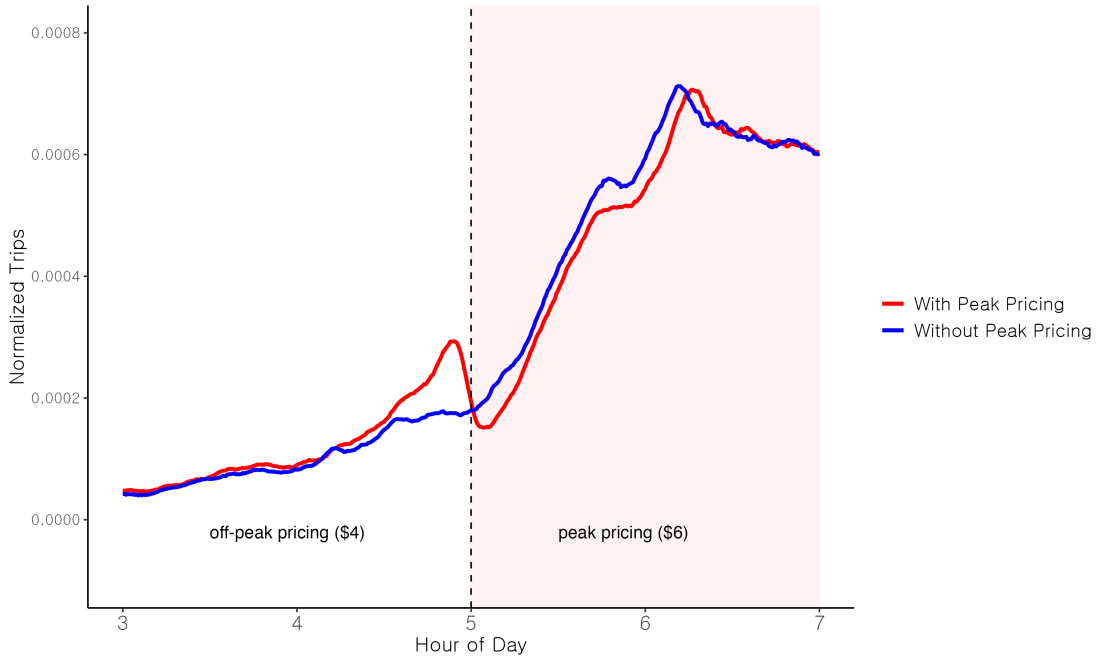

Figure 4: This figure plots the difference in the number of trips in the 6 months before (blue) vs the 6 months after (red) the imposition of peak-hour pricing on the Bay Bridge on July 1, 2010. To facilitate comparison, the number of trips at each time of day is normalized (divided by the total number 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 A6 and A7 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.

Appendix B contains figures that examine the persistence of bunching behavior and the role of tax salience in determining bunching. Figure A8 shows that the bunching behavior is more extreme for drivers who pay in cash than it is for drivers who pay electronic tolls, corroborating findings by Finkelstein (2009). The scheduling cost estimates in Table 4 and elsewhere in this paper reflect the behavior of drivers using electronic toll systems, as this is the technology that would be used in many of the world's proposed cordons. The difference in cash vs. non-cash responses to time-of-day toll systems suggests that factors that increase the salience of electronic tags (e.g., variable message signs displaying cordon costs) may lead to larger temporal adjustment. Figure A7 compares bunching behavior at 6 months, 1 year, and 5 years after the beginning of peak hour pricing: the bunches become smaller over time, and the additional density is spread over a larger off-peak time zone at year 5 than it is at six months. Thus, while some drivers may be able to adjust their ideal arrival times in the long run, the parameters that drive bunching

(schedule costs and ideal arrival times) appear stable for a large fraction of road users.

7.4. 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 subsection. 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 estimates 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). Broadly, the implied value of time in Table 3 is similar to, if slightly larger than, similar estimates from the literature.

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 suggests that drivers prefer being early to being late, and both sets of results suggest that early and late costs are lower than the value of travel time on a per-hour basis.[8]

Finally, it is worth highlighting the potential usefulness of a bunching estimator for studying scheduling costs in other settings. A re-examination of existing empirical work shows this type of temporal bunching is common. For example, bunches around notches in congestion prices are visible in time-of-day plots of traffic in Milan (See figure 4 of Gibson and Carnovale, 2015),

[8] Although the bunching estimator and the discrete choice approach yield scheduling costs that are similar in magnitude, they differ in which direction (early or late) is more costly. This discrepancy arises because these methods use slightly different variation. Scheduling elasticities from the discrete choice model reflect both the price variation induced by the peak-hour tolling period as well as differences in travel times across hours of day.

Bergen (See Figure 6 of Isaksen and Johansen, 2021), Stockholm (See Figure 2 of Kristoffersson, 2013, and London (see Figure 3.8 of Transport for London, 2008). Many bridges and toll roads in the US have similar cutoffs, including roads in Boston, Washington D.C., Seattle, and New York City.

8. Second-Best Optimal Cordon Prices

In this section, I use the discrete choice model estimated in Section 7 together with the tax framework from Section 2 to calculate optimal cordon prices. I first demonstrate this procedure using San Francisco's proposed cordon, and then consider cordon proposed zones in Los Angeles and New York.

At a high level, calculating optimal cordon prices in any city takes four steps: First, I use travel survey data (e.g., the National Household Travel Survey) to identify a representative sample of trips that pass through a city's proposed cordon. Second, I 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, I use the discrete choice model estimated in Section 7 to calculate substitution elasticities between different trips available to drivers. And fourth, I apply the optimal tax formula outlined in Section 2 to the ingredients from steps 1-3.

I refer to each congestion zone as a "proposed" cordon zone. In light of the recent pause in New York's Congestion Pricing Program, it is worth briefly discussing the status of each of these proposed congestion zones. In each city, I take the cordon zone boundary directly from the local transportation agencies planning documents. There are, however, substantial differences across cities in the likelihood that congestion pricing will be implemented in the near future. Although New York has paused its pricing program, the possibility that congestion pricing will come to New York is very real: The policy has federal approval, the infrastructure is in place, and several groups have filed lawsuits against the MTA claiming that the current pause violates the 2019 State law that charges the agency with establishing congestion pricing.

San Francisco's Downtown Congestion Pricing Study was paused in 2022 due to changing traffic and urban conditions in the city following the COVID-19 pandemic. As of 2024, the San Francisco Country Transportation Authority website suggests that the ability to restart the congestion pricing planning process in San Francisco rests with the Transportation Authority Board, and that any potential start date is still at least 5 years out. Of the three cities, Los Angeles is the city furthest from implementing congestion pricing. LA Metro has been studying congestion pricing and receiving public comment since at least 2019, but as of writing, there

is not an official congestion pricing study timeline, nor do the proposed plans explicitly discuss how expensive congestion tolls would be under a hypothetical policy.

8.1. San Francisco's Proposed Cordon Zone

Figure 5 shows a map of the proposed cordon zone in San Francisco and Table A1 the proposed tolling schedule. Both the Figure and the Table reflect the San Francisco Country Transportation Authority's 2021 report on downtown congestion pricing (San Francisco County Traffic Authority, 2021).

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 K, 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 characterization of optimal corrective taxation under preferences for redistribution, see Allcott, Lockwood, and Taubinsky (2019).

Figure 5 — San Francisco's Proposed Congestion Pricing Zone

Figure 5: San Francisco's proposed cordon pricing scheme (San Francisco County Traffic Authority, 2021). Under the 2021 plan outlined by the San Francisco County Transit Authority, trips that enter the cordon would be charged during peak hours 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 Travel 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-On to build a representative dataset of trips that cross San Francisco's proposed cordon zone.[9] 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*

[9]Note 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.

API. The resulting dataset contains 1,891 trips 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 A5.

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 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 A5.

The result of this data collation is a set of trip endpoints for the San Francisco area, where drivers can choose over *route* $\in$ {cordon, non-cordon} and *time of day* $\in \{4.0, 4.2, \ldots, 22.0\}$, 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.

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 A4, 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 6.

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 assigning 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 6.

FIGURE 6 — EXTERNAL COSTS FOR TRIPS CROSSING SAN FRANCISCO'S PROPOSED CORDON

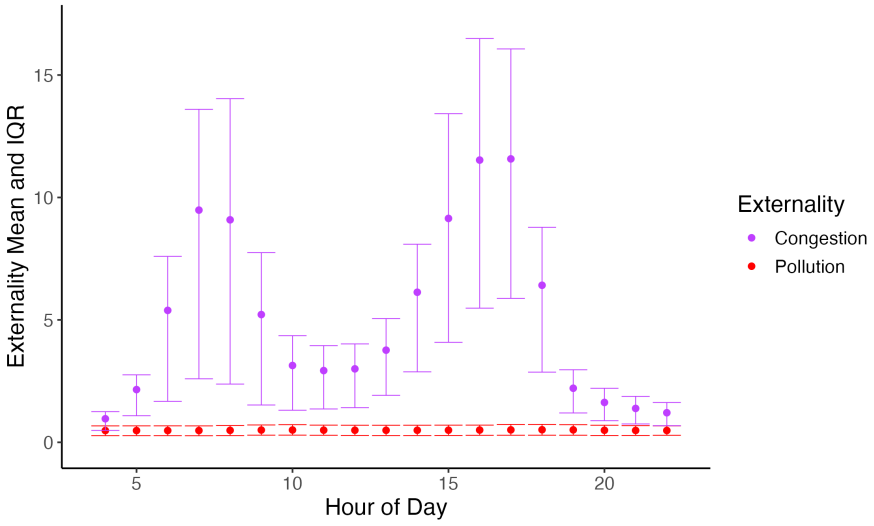

Figure 6: This Figure plots pollution (red) and congestion (purple) externalities by hour for trips in the 2017 National Household Travel Survey (NHTS) with suggested routes that pass through San Francisco's proposed cordon zone. The mean externality within any given hour is represented by a dot; the box spans the 25^{th} to 75^{th} externality percentile, and the bars span the 5^{th} to 95^{th} 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 with 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.

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 7 requires *leakage shares* between trips: $\frac{dh_k}{dp_j} / \frac{dh_j}{dp_j}$. Recall that if j and k are trips (defined as a specific route $\in$ {cordon, non-cordon} at a specific hour of day $\in$ {4.0, 4.2, ..., 22.0}) the leakage share between trip k and trip j represents the share of the reduction in usage of trip k that shifts to trip j as a result 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.

Using Equation 12, leakage shares are implied directly from parameters of the multinomial logit regression estimated in Section 7.

8.5. Optimal Prices

Figure 7 plots three lines relevant for understanding optimal cordon prices. The blue 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 purple line shows these externalities re-weighted as per Diamond (1973) to account for the correlation between the price-responsiveness of trips and idiosyncratic trip-level externalities, as reported in Table 3. Finally, the red (dashed) 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 $5.90 during morning peak hours, and $8.50 during evening peak hours. These second-best optimal prices are calculated using Equation 7, and take into account both the correlation between externalities and elasticities, as well as the substitution to unpriced alternatives in time and space.

FIGURE 7 — SECOND-BEST OPTIMAL CORDON PRICES

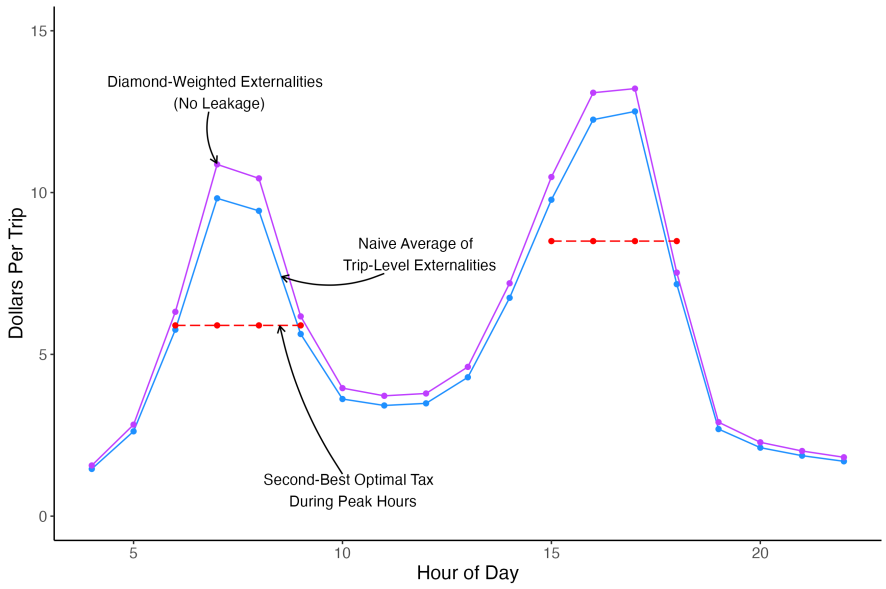

Figure 7: This figure plots three prices relevant for understanding optimal second-best cordon tolls. The blue line plots the average externality (pollution and congestion) for trips that cross San Francisco's cordon by hour of day, estimated using data from the 2017 NHTS (see Section 8.3). The purple line plots externalities re-weighted to account for the correlation between trip-level externalities and trip-level elasticities, as per Diamond (1973). The red (dashed) 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 7, and take 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 7 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 A3 and A15, 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 for the second-best toll of $4.29 and $5.90 for the morning and evening peak hours, respectively.

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

Figure 7 shows that because tolls would incentivize drivers to substitute to routes that avoid the cordon zone, where they would still cause congestion and pollution, the optimal peak-hour cordon prices are below the marginal social damages associated with the average vehicle trip using the cordon zone. In this subsection, I estimate counterfactual driving behavior under a number of tax scenarios to understand the extent to which the imperfections in cordon pricing policies undermine the congestion, pollution, and welfare gains engendered by road pricing policies. These three scenarios are:

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 7. Trips that pass through the cordon area are charged $5.90 during morning peak hours, and $8.50 during evening peak hours (see Figure 7).

I summarize these results in Column 1 of Table 6. Two themes emerge. First, on all three outcome measures — welfare, congestion externalities, and pollution externalities — second-best optimal peak-hour pricing more closely resembles the status quo than the first-best policy.

Second, the welfare results are largely driven by reductions in congestion externalities, as these externalities tend to be 2 to 10 times larger than pollution externalities.

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 8). 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 8 — PROPOSED CORDONS IN NEW YORK AND LOS ANGELES

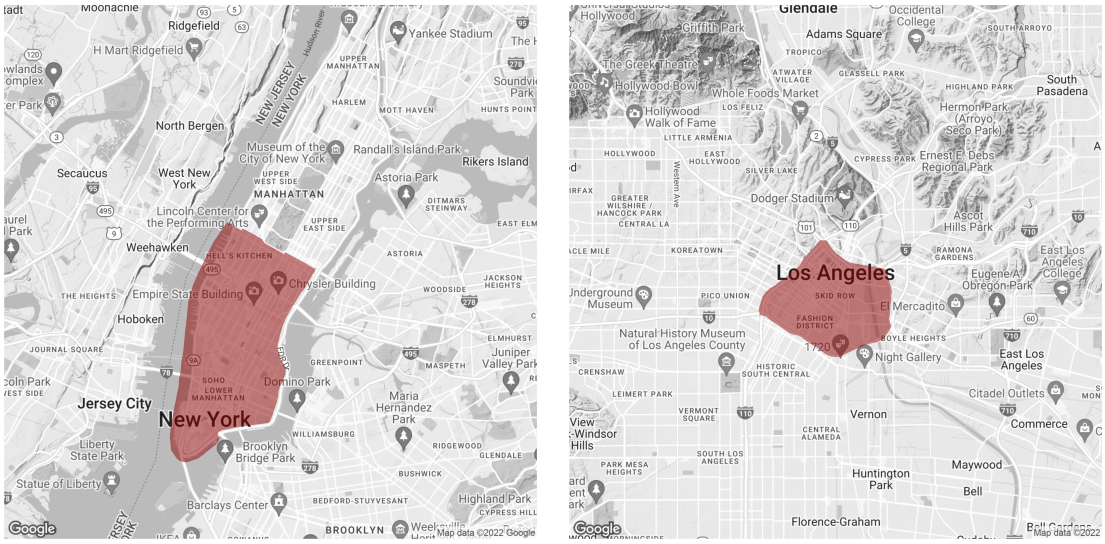

Figure 8: Proposed cordon pricing schemes in New York (Regional Plan Association, 2021) and Los Angeles (Southern California Association of Governments, 2019).

As outlined in Section 2, calculating the second-best optimal cordon prices requires information about the marginal damages of trips that cross through a cordon zone, as well as information about the elasticity and substitutability of these trips. For each of the above cities, I follow the same general template as in San Francisco (see Sections 8.2 through 8.4): First, I use survey data[10] 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 data[11] from city roads to assign congestion externalities to those trips. Third, I calculate substitution parameters between those trips.

[10]The NHTS does not report detailed trip start and end locations for states that are not part of the NHTS Add-On program. The trip-level data for New York come from the 2018 NY Citywide Mobility Survey.

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

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 3). These results are reported in Tables 5 and 6.

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). These proxies for demand primitives are broadly similar across the three cities I examine in this paper. In Appendix J, I document substitution to public transit in response to the increase in Bay Area bridge tolls, and discuss how optimal cordon prices may differ based on the availability of public transportation options. While estimates from a regression discontinuity performed on data from the Bay Area Rapid Transit (BART) system suggest that transit ridership increased after July 2010 (see Table A3), the implied magnitude of mode shifting is small: These estimates suggest that only 6% of drivers who chose not to drive in response to the higher toll prices substituted those trips with BART. In Table A3, I test whether access to public transit impacts the price responsiveness of Bay Area drivers. Point estimates suggest that drivers who live in zip codes near transit stations may be modestly more price-responsive than those who live far away from transit stations, but this difference is not statistically significant. Broadly, the public transit ridership patterns in the Bay Area imply that while some drivers do shift to public transit when the price of their commuting trips increases, these shifts are relatively small, even in transit-rich areas.

Table 5 — Comparing Second-Best Cordon Prices to Social Damages

	Value ($)		
	San Francisco	Los Angeles	New York
Second-Best Price, AM Peak (6-10)	5.90	10.29	11.26
Second-Best Price, PM Peak (3-7)	8.50	11.93	16.31
Average Marginal Damages, AM Peak (6-10)	7.66	14.80	17.60
Average Marginal Damages, PM Peak (3-7)	10.43	17.10	25.03

Table 5: 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 7 — they reflect both heterogeneity in trip-level externalities and leakage in time and space.

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

	Performance Relative to the First-Best (%)		
Outcome	San Francisco	Los Angeles	New York
Welfare Gain	0.411	0.102	0.277
Congestion	0.379	0.11	0.251
Pollution	0.406	0.088	0.377

Table 6: 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 three outcomes of interest are congestion externalities, pollution externalities, and total welfare (the utility of drivers, in dollars, less total externalities). The results in this table reflect the simulated choices of 1.18 million drivers in San Francisco, 953,000 drivers in LA, and 1.43 million drivers in New York. The number of drivers used in each simulation reflects reports from the respective cities. See page 27 of San Francisco County Transportation Authority (2020) for San Francisco, page 41 of Southern California Association of Governments (2019) for Los Angeles, and Table 4A-1 of Metropolitan Transportation Authority (2023). 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 nested logit model shown in Column 3 of Table 3 to the driver choice sets constructed using transportation survey data (see Section 8).

8.8. Hourly Cordon Pricing

Tables 5, and 6 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 13 taxable "goods," where each good covers all cordon trips for a given hour of day $\in \{6, 7, ..., 18\}$.

Table 7 displays 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 improvements, with welfare gains in each city roughly doubling with respect to the gains under uniform peak-hour pricing. In each city, however, a cordon zone with second-best hourly prices would still leave a substantial portion (roughly 20 to 40%) of the possible welfare gains unrealized due to remaining issues of spatial leakage and imprecise pricing.

Table 7 — Back of the Envelope Welfare Gains From Cordon Pricing

	Annual Welfare Gain Relative to Status Quo ($ Million)		
Policy	San Francisco	Los Angeles	New York
First-Best	339	689	1,671
Second-Best (Peak Only)	140	70	463
Second-Best (Fixed Hourly)	272	432	977

Table 7: 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 5); 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 7 — they reflect both heterogeneity in trip-level externalities and leakage in time and space. The figures in this table reflect simulated choices using the nested logit model shown in Column 3 of Table 3. The results in this table reflect the simulated choices of 1.18 million drivers in San Francisco, 953,000 drivers in LA, and 1.43 million drivers in New York. The number of drivers used in each simulation reflects reports from the respective cities. See page 27 of San Francisco County Transportation Authority (2020) for San Francisco, page 41 of Southern California Association of Governments (2019) for Los Angeles, and Table 4A-1 of Metropolitan Transportation Authority (2023).

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. Whether these imperfections lead to optimal prices that are above or below trip-level damages depends on the degree and sign of the correlation between price-responsiveness and idiosyncratic externalities, and the strength of the leakage effect.

The results from Figure 7 show that the leakage effect depresses second-best peak-hour congestion prices. In San Francisco, optimal prices are $5.90 for the morning peak period and $8.50 for the evening peak period — significantly below the social cost for trips that pass through the cordon at those times.

My findings suggest that if the determinants of driver decisions (price responsiveness, value of travel time, schedule flexibility) are similar across large US cities, then optimal cordon prices

are also below the average of social damages generated by downtown trips in New York and Los Angeles. Table 5 shows that in New York, for example, the second-best optimal cordon prices are about $11.26 and $16.31 for morning and evening peaks, respectively, which are below the average social damages associated with cordon trips in each of those periods ($17.60 and $25.03, respectively). In Los Angeles, the optimal morning and evening peak prices are $10.29 and $11.93, compared to average social damages of $14.80 and $17.10.

Cordon zones charging the second-best prices described in Table 5 would generate significant welfare gains for commuters and city residents in New York ($463 million annually) and San Francisco ($140 million annually), but modest benefits in Los Angeles ($70 million annually). 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 half to one percent of city budgets in New York and San Francisco.

Despite these welfare gains, the results in Section 8 suggest that the blunt nature of cordon pricing limits their effectiveness relative to an ideal policy. In San Francisco, optimal peak-hour cordons achieve 41% of the welfare gains that would be realized under a first-best policy. This ratio is even lower in New York (28%) and Los Angeles (10%). Notably, the welfare gains of peak-hour pricing policies largely track reductions in congestion externalities. This reflects the fact that on a per-trip basis, the estimated social damages associated with congestion tend to be much larger than the social damages associated with pollution (see Figure 6). This feature may not hold in metro areas outside of the US, which tend to have vehicle fleets with higher per-mile local pollution rates.

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 6) provides sizeable welfare gains: $132 million in San Francisco, $362 million in Los Angeles, and $514 million in New York. In each city, however, this flexible pricing strategy still leaves a significant portion of the first-best welfare gains unrealized.

This paper takes as given the spatial layout of cordon zones. The difference in performance between the proposed zone in Los Angeles and those in New York and San Francisco, however, highlights the central role of leakage in the effectiveness of cordon zones. Before bringing in any information about trips or externalities, the proposed zone in Los Angeles appears more likely to invite spatial leakage. Because downtown LA lacks the bounding bodies of water seen in downtown New York and San Francisco, there are many busy roadways just outside the boundary of LA's proposed cordon zone. Indeed, the trip-level data suggest that just under 57% of trips that pass through LA's proposed zone have viable detours that avoid the zone — higher than in New York (45%) or San Francisco (19%). These trips are also similar in damages to the

cordon zone trips: On average, detour trips in LA have social costs that are 95% as high as the average cordon zone trip. In New York, this ratio is 41%. In San Francisco, detour trips actually have *higher* social damages than the average cordon trip, but as mentioned above, the majority of trips do not have a viable detour.

To summarize, the proposed cordon zone in Los Angeles is both easy to avoid for drivers *and* the substitute routes are associated with high social costs. As a result, even second-best optimal prices yield relatively small welfare gains.

What lesson does the above comparison hold for policy? Policymakers may want to set congestion zone 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 two- 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 peak-hour 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 is larger than the heterogeneity effect (see Diamond, 1973), which also depresses second-best optimal prices.

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 $140 million worth of welfare gains, this policy would fall short of the first-best policy by $132 million annually. This foregone welfare is significant: $132 million is roughly 1% of the City of San Francisco's 2020-2021 Budget of $13.7 billion. The predicted performance of proposed cordons in New York and Los Angeles are qualitatively similar. The role of leakage in this setting suggests that there may be large gains from designing cordon zones to preempt spatial leakage and that certain

cities may see larger welfare gains due to cordon pricing because of idiosyncratic city geography that makes spatial trip substitution more difficult.

Lastly, this paper contributes to the literature in 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, suggests 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.

Bertoli, Simone, J Fernández-Huertas Moraga, and Francesc Ortega. 2013. "Crossing the border: Self-selection, earnings and individual migration decisions." *Journal of Development Economics* 101:75–91.

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.

Conlon, Christopher and Julie Holland Mortimer. 2021. "Empirical properties of diversion ratios." *The RAND Journal of Economics* 52 (4):693–726.

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.

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.

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.

Fleissig, Adrian R. 2021. "Estimating elasticities of substitution for sin goods." *Applied Economics* 53 (30):3549–3561.

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

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 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.

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.

Isaksen, Elisabeth Thuestad and Bjørn Gjerde Johansen. 2021. "Congestion pricing, air pollution, and individual-level behavioral responses." *Available at SSRN 3832230* .

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

Kovach, Matthew and Gerelt Tserenjigmid. 2022. "The focal Luce model." *American Economic Journal: Microeconomics* 14 (3):378–413.

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

Kristoffersson, Ida. 2013. "Impacts of time-varying cordon pricing: Validation and application of mesoscopic model for Stockholm." *Transport Policy* 28:51–60.

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.

Metropolitan Transportation Authority. 2023. "Regional Transportation Effects and Modeling." Tech. rep., Central Business District (CBD) Tolling Program Environmental Assessment.

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.

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.

San Francisco County Transportation Authority. 2020. "San Francisco Mobility, Access, and Pricing Study." Tech. rep., San Francisco County Transportation Authority.

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

Selmoune, Aya, Qixiu Cheng, Lumeng Wang, and Zhiyuan Liu. 2020. "Influencing factors in congestion pricing acceptability: a literature review." *Journal of Advanced Transportation* 2020 (1):4242964.

Sheu, Gloria. 2014. "Price, quality, and variety: Measuring the gains from trade in differentiated products." *American Economic Journal: Applied Economics* 6 (4):66–89.

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.

Southern California Association of Governments. 2019. "Mobility Go Zone Pricing Feasibility Study." Tech. rep., Southern California Association of Governments.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.

Transport for London. 2008. "Central London congestion charging impacts monitoring sixth annual report, July 2008." Tech. rep., Transport for London.

Tseng, Yin Yen, Barry Ubbels, and Erik Verhoef. 2005. "Value of time, schedule delay and reliability." In *ERSA conference papers*.

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.