CAN TOLLING HELP EVERYONE? ESTIMATING THE AGGREGATE AND DISTRIBUTIONAL CONSEQUENCES OF CONGESTION PRICING*

JONATHAN D. HALL
UNIVERSITY OF TORONTO

January 15, 2020

ABSTRACT. Economists have long advocated road pricing as an efficiency-enhancing solution to traffic congestion, yet it has rarely been implemented because it is thought to create losers as well as winners. In theory, a judiciously designed toll applied to a portion of the lanes of a highway can generate a Pareto improvement, even before using the toll revenue. This paper explores the practical relevance of this theoretical possibility by using survey and travel time data, combined with a structural model of traffic congestion, to estimate the joint distribution of agent preferences over three dimensions—value of time, schedule inflexibility, and desired arrival time—and evaluate the effects of adding optimal time-varying tolls. I find that adding tolls on half of the lanes of a highway yields a Pareto improvement. Further, the social welfare gains from doing so are substantial—up to $1,740 per road user per year.

JEL Classification: D62, H41, R41, R48.

*I am especially grateful for the guidance and support that Gary Becker and Eric Budish have given me. I am also grateful for helpful feedback from Claudio Michelacci, two anonymous referees, Richard Arnott, Ian Fillmore, Mogens Fosgerau, Edward Glaeser, Brent Hickman, Chelsea Hall, William Hubbard, Kory Kroft, Ethan Lieber, Robin Lindsey, Robert McMillan, Peter Morrow, John Panzar, Devin Pope, Mark Phillips, Allen Sanderson, Ken Small, Chad Syverson, George Tolley, Vincent van den Berg, Jos van Ommeren, Clifford Winston, and Glen Weyl, as well as seminar audiences at the University of Chicago, Northwestern University, University of Toronto, Brigham Young University, Clemson University, Technical University of Denmark, Tinbergen Institute, HEC Montreal, RSAI, Kumho-Nectar, World Bank Conference on Transport and ICT, and the NBER Summer Institute. All remaining errors are my own. *Email:* jonathan.hall@utoronto.ca.

This is a pre-copyedited, author-produced version of an article accepted for publication in the Journal of the European Economic Association following peer review. The version of record is available online at: http://dx.doi.org/10.1093/jeea/jvz082.

1. Introduction

In the ninety-eight years since Pigou (1920) introduced the idea that adding tolls could alleviate traffic congestion, horse-drawn carts have given way to automobiles and congestion has grown to consume 42 hours per urban commuter in the United States annually (Schrank et al. 2015). Commuting is self-reported to be the least pleasant activity in the day (Kahneman et al. 2004), and all this extra time spent in traffic wastes fuel, causes additional pollution, and leads to serious health problems.[1] Because of the seriousness of these costs, addressing traffic congestion has become a leading issue for cities around the world.[2]

Notwithstanding these pressing concerns, governments remain hesitant to turn to tolling to alleviate traffic congestion, in large part because of the received wisdom that adding tolls hurts many, if not most, road users.[3] As one voter put it, "Turkeys don't vote for Christmas–and motorists won't vote for more taxes to drive."[4]

The distributional consequence of tolling have been the focus of a long literature.[5] I build on this literature in related theoretical work that shows it is possible for a carefully designed toll applied to a portion of the lanes of the highway to generate a Pareto improvement, even before the revenue is spent (Hall 2018). Generating a Pareto improvement before using the revenue is valuable because while in principle there exists a set of transfers such that congestion pricing helps

[1]Schrank et al. (2015) reports that congestion wastes 3.1 billion gallons of fuel per year. Currie and Walker (2011) estimates that the pollution due to traffic congestion is responsible for up to 8,600 preterm births a year.

[2]Bloomberg Philanthropies (2018) surveys mayors from across the U.S. and find that 41% of them report that traffic congestion is one of their top three problems. Langan et al. (2017) analyze 120 "State of the City" speeches from a balanced sample of U.S. cities by size and geographic area, and find that "roads" is the third most common subtopic, out of a list of 155 possible subtopics, and was a covered in 48% of all speeches. Furthermore, my own investigation of all the "State of the City" speeches from the primary cities of the 30 largest metropolitan areas in the U.S. in 2017 found that 46% of them express concerns about traffic congestion.

[3]For an excellent explanation of this standard result, see Small and Verhoef (2007, pp. 120–127). For examples of papers citing these losses as a major barrier to implementing tolling, see Starkie (1986), Cohen (1987), Giuliano (1992), Arnott et al. (1994), Lave (1994), Santos and Rojey (2004), Sorensen and Taylor (2005), Small et al. (2005), Small and Verhoef (2007), Lindsey and Verhoef (2008), and van den Berg and Verhoef (2011).

[4]Sturcke, James. 2008. "Manchester Says No to Congestion Charging," Guardian, December 12, 2008.

[5]See, for example, Foster (1975), Arnott et al. (1994), Small (1983, 1992), Santos and Rojey (2004), Small et al. (2006), Schweitzer and Taylor (2008), Light (2009), and van den Berg and Verhoef (2011).

everyone, it is difficult to actually implement such transfers.[6] However, whether it is possible to obtain a Pareto improvement without using the revenue depends on the distribution of traveler preferences, and furthermore, being able to price some portion of the lanes does not guarantee it is possible to price an economically meaningful portion of the lanes.

In this paper, I investigate the practical relevance of the theoretical possibility of generating a Pareto improvement by estimating the aggregate and distributional consequences of congestion pricing. I do so by, first, extending the model of Hall (2018), itself a modified version of the bottleneck model of Vickrey (1969) and Arnott et al. (1993), to allow for continuous, rather than discrete, heterogeneity in value of time and schedule inflexibility. Extending the model in this way makes it more empirically relevant, but does not affect the underlying intuition or mechanisms.

Second, I combine the model with survey and travel time data from California State Route 91 to estimate the joint distribution of traveler preferences over three dimensions: value of time, schedule inflexibility, and desired arrival time. This is the first time the distribution of schedule inflexibility has been estimated, as well as the first time this joint distribution has been estimated. Estimating the distribution of inflexibility is important as it is the structural object at the heart of any model involving the scheduling of activities, and my estimates are relevant to a wide variety of questions regarding urban transportation and land use.[7] I estimate the distributions of value of time and desired arrival time from survey data, and then estimate the distribution of schedule inflexibility such that the model-predicted travel times best match observed travel times. Identification of schedule inflexibility (a traveler's willingness to arrive early or late to reduce travel time) comes from observing travelers' choice sets (all feasible combinations of arrival time and travel time), and using the requirement that the marginal rate of substitution is tangent to budget constraint (in this case, the upper bound of the

[6] See Foster (1975) for a short discussion of the difficulty in targeting the transfers. Stiglitz (1998) notes that even when we can design such transfers, they can be difficult to implement for at least two reasons. First, the transfers are transparent and so harder to defend than the implicit transfers entailed by the status quo. Second, the government cannot commit to maintaining the transfers, and so voters may not believe they are better off in the long run.

[7] For example, questions about highway or mass transit capacity, parking policies, mass transit frequency, the effects of tolls on urban spatial structure, and airline competition all depend on schedule inflexibility. For examples of papers dealing with these questions, see Arnott et al. (1993), Fosgerau and de Palma (2013), Silva et al. (2014), Takayama and Kuwahara (2017), and de Palma et al. (2017).

choice set). Thus, as in Larsen and Zhang (2018), agent preferences are identified by the slope of the choice set, evaluated at the choice made by the agent.

Third, I use these estimates and my model to evaluate the aggregate and distributional effects of congestion pricing, finding that the welfare gains from congestion pricing are large. Pricing all the lanes increases social welfare by $2,400 per road user per year, but at the cost of hurting some road users by $2,390 per year. However, pricing just half of the lanes generates a Pareto improvement while still increasing social welfare by $1,740 per road user per year. I extrapolate my results to the rest of the United States by assuming traveler preferences are the same in all cities, and adjusting for the severity of congestion and miles traveled in each city. I find that pricing half the lanes on all urban highways would increase social welfare by over $30 billion per year, or $850 per year for the typical urban highway commuter.[8]

This paper contributes to two literatures: first, the literature estimating how individuals choose when and how to travel. This paper is closely related to Small (1982), who provided the first estimates of average schedule inflexibility, and Small et al. (2005) who combine stated and revealed preference data with a structural choice model to estimate the distribution of value of time.

Second, this paper contributes to the long literature estimating the aggregate and distributional effects of congestion pricing.[9] Within this literature, I build the most closely on an innovative paper by van den Berg and Verhoef (2011), which extends the bottleneck model to allow agent preferences to vary continuously on two dimensions: value of time and schedule inflexibility. They show numerically that for intuitively reasonable parameter values, pricing all the lanes does not always hurt the majority of agents and that it is possible to generate a Pareto improvement by pricing a third of the lanes and forgoing revenue by charging a negative toll off-peak.

I extend van den Berg and Verhoef (2011) in three ways. First, I estimate the estimated joint distribution of agent preferences, second, I allow for an important additional traffic externality (which I explain in Section 2) and third, I allow

[8]The social welfare gains are smaller for the typical urban road user than for those on California State Route 91 because Route 91 is among the most congested highways in America and those who use it have longer-than-average commutes.

[9]Other important papers in this literature include Winston and Langer (2006), Light (2009), and Couture et al. (2018).

for travelers to differ along an additional dimension: desired arrival time. Allowing heterogeneous desired arrival times matters because it is necessary to match empirical travel times using reasonable parameter values.

This paper is also closely related to an important paper by Small et al. (2006), who estimate traveler preferences using revealed and stated preference data and a discrete choice model, and use their estimates to evaluate the aggregate and distributional effects of a variety of congestion pricing policies. They find that a variety of second-best policies can significantly reduce the share of travelers who are worse off. I build on Small et al. (2006) by using a structural model of congestion which explicitly accounts for the interaction between road users over time, and allows road users to choose when to travel. I further extend their work by estimating travelers' preferences over arrival times, both in terms of their ideal arrival time and how costly it is for them to arrive early or late (i.e. schedule inflexibility).

2. Understanding Hypercongestion and How It Affects the Distributional Consequences of Tolling

As this paper builds on Hall (2018), in this section I summarize the intuition for the key results of this paper. The key result is that it is possible for a time-varying toll on a portion of the lanes of a highway to help all road users, even before the revenue is spent and even with realistic heterogeneity. Pricing a portion of the lanes is often called value pricing, and are called HOT lanes (high-occupancy toll lanes) when those carpooling can use them for free.

Central to this result is an important additional traffic externality: not only does each additional vehicle slow others down, but in heavy enough traffic, additional vehicles can create frictions that reduce throughput (the number of trips per unit time). To understand the two externalities better, consider a production possibilities frontier (PPF) representing the trade-off between the number of people traveling (or throughput) and speed, as shown in Figure 1. In this figure, the standard externality moves equilibrium along the PPF from point A to point B, and so away from the point that maximizes social welfare, while the additional externality moves equilibrium off the PPF to point C.

The economics literature has long debated the possibility that too many vehicles on the road could reduce throughput, and Vickrey (1987) even gave this second externality a name: "hypercongestion." However, as the theoretical arguments

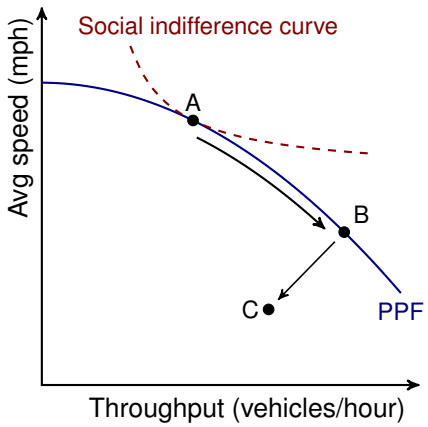

FIGURE 1. Production possibilities frontier illustrating both externalities

for hypercongestion did not hold in dynamic models of highway congestion, the literature cast doubt on its existence.[10]

Recently, however, the transportation engineering literature has identified two causal mechanisms for hypercongestion and provided extensive empirical evidence that it is an important real-world phenomenon. The first mechanism occurs when the queue behind a bottleneck grows long enough that it blocks other traffic.[11] For example, the queue at a busy off-ramp might grow until it spills over on to the mainline of the highway, blocking at least one lane and often blocking multiple lanes. It is this mechanism that makes beltways and ring roads especially prone to crippling congestion, as their circular nature makes it possible for drivers to simultaneously be in each other's way (Vickrey 1969, Daganzo 1996).

The second causal mechanism for hypercongestion is due to frictions that occur when merging at the bottleneck itself. In heavy traffic, it is difficult for vehicles in the lane that is ending to merge, and there is eventually a vehicle that must slow down, or even stop, before merging. This vehicle often engages in what transportation engineers call a "destructive lane change," forcing its way over

[10]The theoretical possibility of hypercongestion comes from the fact that throughput (vehicles/hour) is the product of speed (miles/hour) times density (vehicles/mile), $T = S \times D$, and since speed is decreasing in density, $dS/dD < 0$, then it is possible for throughput to be decreasing in density, $dT/dD = D \cdot dS/dD + S \gtrless 0$ (e.g., Walters 1961, Johnson 1964, De Meza and Gould 1987). The literature correctly responded that hypercongestion is a dynamic phenomenon, and showed that in dynamic models the mathematical relationships above were not enough to generate hypercongestion (e.g., Newell 1988, Evans 1992, Verhoef 1999, 2001, 2005, May et al. 2000, Small and Chu 2003).

[11]A bottleneck can occur at any place where highway capacity decreases, generally due to a reduction in the number of lanes. Perhaps the most common cause of bottlenecks are on-ramps.

and reducing throughput. Throughput falls because the vehicle goes through the bottleneck at a slow speed; equivalently, the vehicle opens up a gap in front of itself, creating a period of time where the bottleneck—the scarce resource on the highway—is not being used.

Hall (2018) summarizes the empirical evidence for both these mechanisms, reporting that queue spillovers reduce throughput by approximately 25%, while destructive lane changes at bottlenecks reduce throughput by 10% on average.

It is by reducing, or even eliminating, this second externality that adding tolls can generate a Pareto improvement. Returning to Figure 1, moving traffic to a point to the northeast of C increases both speeds and throughput, and, when agents are homogeneous, this is enough to conclude everyone is better off, even before using the revenue. Of course, road users are not homogeneous, and allowing for heterogeneity makes it likely that pricing all of the lanes hurts some road users, even when tolling increases throughput. Adding tolls reduces the time costs of traveling while increasing the financial costs. As not all drivers value their time the same, this likely hurts some drivers.

However, it is possible to obtain a Pareto improvement by adding tolls to just a portion of the lanes. Pricing a portion of the lanes increases highway throughput, while preserving the ability of those with low values of time to pay with time, rather than money, to travel on the highway during rush hour. As the benefits of increasing throughput accrue to those in both the priced and free lanes, this helps all road users, even before using the toll revenue.

Whether this intuition holds depends on the distribution of driver preferences. This is because adding tolls can cause drivers who were traveling off-peak to start traveling at the peak, displacing those who used to travel at the peak. This displacement hurts these drivers, making it difficult to generate a Pareto improvement (unless the magnitude of the second externality is implausibly large). Whether this occurs depends on the distribution of driver preferences. Displacement will not be a problem if there are a sufficient number of drivers with high values of time traveling at the peak when the road is free, as these drivers will be the ones to use the priced lanes during the peak.

3. Model

The foundation for my model is the bottleneck model of Vickrey (1969) and Arnott et al. (1993). I build most directly on the model in Hall (2018), which I

extend to allow agent preferences over value of time and schedule inflexibility to be continuously distributed.[12]

3.1. **Congestion Technology.** A single road connects where people live to where they work; this road can be split into two routes, one tolled, the other free.[13] Let λ_{toll} and λ_{free} denote the fraction of capacity devoted to each route, where $\lambda_{\text{toll}} + \lambda_{\text{free}} = 1$.[14] Travel along this road is uncongested, except for a single bottleneck through which at most s^* vehicles can pass per unit time. Let $\rho_r(t)$ denote the departure rate of vehicles from home using route r at time t. When this departure rate exceeds the routes capacity, $\lambda_r \cdot s^*$, a queue develops. Once the queue is more than ϵ vehicles long the throughput of the bottleneck for that route falls to $\lambda_r \cdot s$, where $s \leq s^*$. Therefore, queue length, Q_r, measured as the number of vehicles in the queue, evolves according to

$$\frac{\partial Q_r(t)}{\partial t} = \begin{cases} 0 & \text{if } Q_r(t) = 0 \text{ and } \rho_r(t) \leq \lambda_r \cdot s^*, \\ \rho_r(t) - \lambda_r \cdot s^* & \text{if } Q_r(t) \leq \epsilon \text{ and } \rho_r(t) > \lambda_r \cdot s^*, \\ \rho_r(t) - \lambda_r \cdot s & \text{if } Q_r(t) > \epsilon. \end{cases}$$

[12] van den Berg and Verhoef (2011) was the first, and, until now, only, paper to allow continuously distributed values of time and schedule inflexibility in the bottleneck model.

[13] In Appendix C.4 and C.5 I discuss how the assumptions that everyone has the same commute, and that it is possible to price the entire commute (i.e., there are not local roads at the start or end of the trip) affect my results.

[14] Implicit in this is the assumption it is costless to split the road into two routes. In reality some separation between the priced and unpriced lanes is required. The Federal Highway Administration recommends a three to four foot buffer when a pylon barrier is used (Perez and Sciara 2003, 39-40) and on I-394 in Minnesota there is a two foot buffer without any barrier (Halvorson and Buckeye 2006, 246). As federal standards call for twelve foot lanes on interstates (AASHTO 2005, 3), splitting the road into two routes could cost as much as a third of a lane. This space can come from narrowing the existing lanes at the cost of reducing the design speed of the highway or the highway could be widened by a few feet. In addition, in reality we are constrained to pricing an integer number of lanes. This matters when pricing two-lane highways, but is less of an issue on the typical wide urban highway.
That said, there are two reasons splitting the road into two routes may increase capacity. First, while the literature on throughput drops has not explicitly considered how the magnitude of the throughput drop is related to the number of lanes, Zhang and Levinson (2004) estimates the size of the throughput drop for 27 sites in the Twin Cities in Minnesota, reporting the number of lines and estimated throughput drop for each site. In this data, the magnitude of the throughput drop is *increasing* in the number of lanes. However, this relationship is (marginally) not statistically significant (p-value of 0.07). Accounting for the fact that the estimated throughput drop is a generated regressor would further reduce the statistical significance.
Second, Menendez and Daganzo (2007) and Cassidy et al. (2010) show that splitting a road into two routes by adding an HOV lane *increases* flow in the adjacent general purpose lanes by reducing the number of lane changes.

I then simplify by taking the limit as $\epsilon \to 0$, so throughput on a congested route is constant.

Travel time on route r for an agent arriving at t is

$$T_r(t) = T^f + T_r^v(t),$$

where T^f is fixed travel time—the amount of time it takes to travel the road absent any congestion—and $T_r^v(t)$ is variable travel time for route r. For simplicity, and without loss of generality, let $T^f = 0$. Throughout the rest of this paper when I discuss travel time I am only referring to the variable congestion-related travel time.

In Appendix A I show the production possibilities frontier (PPF) of the bottleneck model is similar to one I estimate using high-frequency data from an isolated bottleneck in San Diego, as well as being like common models used in the transportation engineering literature.

3.2. **Agent Preferences.** Agents choose when to arrive at work and which route to take to minimize the cost of traveling. Agents dislike three aspects of traveling: travel time, tolls, and schedule delay—that is, arriving earlier or later than desired. These costs combine to form the trip cost; the trip cost of arriving at time t on route r for an agent of type i with desired arrival time t^* is

$$p(t, r; i, t^*) = \alpha_i T_r(t) + \tau_r(t) + D_i(t^* - t) \tag{1}$$

where α is the cost per unit time traveling (i.e., the agent's value of time), $\tau_r(t)$ is the toll charged on route r for arriving at time t, and D_i is type i's schedule delay cost function. Schedule delay costs are piecewise linear,

$$D_i(t^* - t) = (t^* - t) \begin{cases} \beta_i & t \leq t^* \\ -\gamma_i & t > t^* \end{cases}$$

where β is the cost per unit time early to work, and γ is the cost per unit time late to work. Each of these parameters represents how much an agent is willing to pay in money to reduce travel time or schedule delay by one unit of time.

It is worth highlighting that arriving later than desired does not necessarily imply arriving literally late. As a simple example, consider the problem of scheduling a medical appointment. In a world free of traffic congestion, someone might prefer to schedule it first thing in the morning. But given how bad traffic will be then, he instead schedules the appointment for later in the day. He arrives

on-time for his appointment, but since it was scheduled for later than desired, still has schedule delay costs. Similarly arriving earlier than desired does not mean arriving literally early.

Let $\beta_i < \alpha_i$ for all i. This means that agents would rather wait for work to start at the office than wait in traffic and is needed to prevent the departure rate from being infinite.

To simplify, let $\gamma_i = \xi\beta_i$ for all i, with ξ a constant scalar. This means that those who dislike being early also dislike being late, while those who do not mind being early similarly do not mind being late.[15]

Agents can differ in their value of time, schedule delay costs, and desired arrival time. Each of these are continuously distributed. I consider agents to be of the same *type* if they have the same value-of-time and schedule delay costs. Let $\mathcal{G}$ denote the set of types.

The primary source of heterogeneity in agents' value of time is variation in their income, and so if $\alpha_i > \alpha_j$ then type i is *richer* than type j. While there are other sources of heterogeneity in agents' value of time,[16] by using α as a proxy for income we can directly discuss the primary concern with congestion pricing: that it helps the rich and hurts everyone else.

The ratios β/α and γ/α are an agent's willingness to pay in travel time to reduce schedule delay (early and late respectively) by one unit of time and provide a measure of how inflexible his schedule is, so let $\delta_i = \beta_i/\alpha_i$ be type i's *inflexibility*. A major source of heterogeneity in agents' flexibility arises from differences in occupation, as the opportunity cost of time early or late is different for those with different types of jobs. If a shift worker is late he generally faces penalties and when he is early he passes the time talking with co-workers. Since there is not much difference for the shift worker between spending time traveling or being at work early, his δ is close to one (the largest possible δ). Similarly, due to the penalty when late, $\xi\delta = \gamma/\alpha$ is large. In contrast, an academic can start working

[15]Relaxing this assumption would only affect my results if there are agents who switch from arriving early to arriving late, or vice-versa, when tolls are added to the road. Because of this assumption, my estimator for marginal distribution of β/α (and the distribution of γ/α as it is a transformation of the distribution of β/α) uses information about both early and late arrivals. In Section 5 I also fit a version of the model which relaxes this assumption, among others, and find that these assumptions have a fairly trivial effect on how well I can match the data and on those parameter estimates the relaxed version of the model can recover.

[16]For example, trip purpose, distance and mode (Abrantes and Wardman 2011).

whenever he gets to the office and so has a very low marginal disutility from being early or late and so his δ is closer to zero.

Another major source of heterogeneity in agents' flexibility comes from their personal lives. Arriving at work earlier means waking up earlier, and so going to bed earlier, which may interfere with social plans. Similarly, arriving at work later likely implies working later to make up for lost time. Furthermore, arriving early or late to work may interfere with family responsibilities.

Within each type, agents' desired arrival times are uniformly distributed over $[t_s, t_e]$. Having a continuous distribution of desired arrival times allows a positive measure set of agents to arrive on-time, and thus allows agents to be inframarginal with respect to the choice of when to arrive, meaning that they will not change their chosen arrival time in response to a small increase in the cost of that arrival time, holding all else constant. I call these agents *inframarginal*, suppressing the specification of which dimension of choice they are inframarginal on.

Allowing for inframarginal agents is essential for taking the model to the data. The standard bottleneck model cannot reproduce realistic travel times using reasonable parameter values, however, allowing for inframarginal agents solves this problem (Hall 2019).

Allowing for inframarginal agents also affects the qualitative results. Inframarginal agents value their arrival time above their next best alternative at existing prices. Adding tolls alters the prices, requiring some inframarginal agents to either pay a price reflecting their true valuation or switch to a less desirable alternative. This change hurts some of these agents and is a significant barrier to obtaining a Pareto improvement.

Furthermore, having a continuum of desired arrival times is more reasonable than it initially sounds. While it may seem more natural to assume an agent's desired arrival time falls into some discrete set, such as $\{7{:}00, 7{:}30, \ldots, 9{:}00\}$, what matters is when agents want to arrive at the end of the highway, not when they want to arrive at work. Because the distribution of distances between the end of the highway and work is continuous, the distribution of desired arrival times at the end of the highway is also continuous.

Assuming the distribution of desired arrival times is uniformly distributed, rather than using a different continuous distribution, keeps the model analytically tractable despite having a continuum of types.[17]

[17] In Appendix C.6, I provide evidence that this is a reasonable approximation to the truth.

Let n_i denote the density of agents of type i who desire to arrive at a given time in $[t_s, t_e]$. I normalize the total mass of agents to one. Furthermore, $\sum n_i$ is assumed to exceed the road's capacity (s^*), so it is impossible for all agents to arrive at their desired arrival time; thus, some need to arrive early or late.

The mass of agents of each type who use the road is independent of the trip cost, that is, demand for travel along this road is perfectly inelastic. Were demand not perfectly inelastic then the distribution of desired arrival times would no longer be uniform once tolls were added to the highway and different types saw their trip costs change by different amounts. By assuming perfectly inelastic demand, I maintain the benefits of having uniformly-distributed desired arrival times. This assumption fits my empirical context: California State Route 91, a highway through a mountain pass between Riverside County and Orange County. Commuters do not have a reasonable alternative to taking SR-91 and public transit accounts for less than 1% of the trips through the pass (Sullivan and Burris 2006, 192); thus I do not need to worry about agents switching to different roads or modes and the only choice I am missing is the choice of whether to travel.

By having perfectly inelastic demand I rule out one way pricing can hurt the poor: because congestion pricing lowers the cost for richer agents it induces more rich agents to travel. This counteracts some of the benefit to existing agents of increasing throughput. If demand for trips by the rich is sufficiently elastic, it is even possible rush hour is longer once congestion pricing is implemented. In a previous version of this paper I had elastic demand (with homogeneous desired arrival times) and the elasticity of demand only had minor effects on the results. That said, there is evidence that the long run demand for travel is perfectly elastic (Duranton and Turner 2011). If demand is perfectly elastic for all types, then each types' demand curve is horizontal. When demand curves are horizontal, the equilibrium cost of travel is the same regardless of how supply changes. Since the cost of travel for each type is constant, pricing neither helps nor hurts any agents before the revenue is used.[18]

3.3. **Definition of Equilibrium.** Let $\{r, t\} = \sigma(i, t^*)$ be the strategy of an agent of type i with desired arrival time t^*; $\sigma : \mathcal{G} \times [t_s, t_e] \to \{\text{free, toll}\} \times [0, 24]$.

[18]This holds regardless of how pricing affects throughput. Pricing will change the quantities of each type using the road.

The relevant equilibrium concept is that of a perfect information, pure strategy Nash equilibrium, in which no agent can reduce his trip cost by changing his arrival time or route choice.

I show that an equilibrium exists by construction, and show that equilibrium trip prices, travel times, and tolls are unique in Appendix B.

4. Finding the Equilibrium

In this section I give an overview of how to solve for the equilibrium trip price. Allowing agent preferences to be continuously distributed along three dimensions (value of time, schedule inflexibility, and desired arrival times) makes the model more empirically relevant, however, it also introduces complications into solving for the equilibrium. To save space, the details of doing so are relegated to Appendix B.

4.1. **Equilibrium When Road Completely Free or Priced.**

To solve for equilibrium when all of the lanes are free, I first show that, conditional on which agents arrive early, on-time, and late, I can order agents arrival times using their inflexibility (more flexible agents arrive very early or very late). This gives me an assignment of agents to arrival times, and given this, I can back out from agent preferences what travel times are, and so can calculate each agent's travel time and schedule delay. This lets me write down each agent's trip cost as a function of which agents arrive before their desired arrival time. I then use equilibrium conditions to write down a functional equation defining which agents actually arrive early, on-time, and late, and solve this. Solving this functional equation is non-trivial, in the main difficulty in solving for equilibrium.

Solving for equilibrium when all the lanes are tolled is very similar, except that in this case the ordering depends on agents' β rather than agent's inflexibility (δ). In solving, I assume that the toll is zero when the road is uncongested.[19]

I derive the following closed form solutions for equilibrium trip costs on a completely free or priced route:

$$\bar{p}_{\text{free}}(\alpha, \delta, t^*) = \frac{\tilde{\xi}}{1+\tilde{\xi}} \frac{1}{s} \left[\alpha \int_0^1 \min\{\delta', \delta, \hat{\delta}\} n_\delta(\delta')\, d\delta' \right] \tag{2}$$

[19] Allowing the toll to be negative is an effective way to "spend" the toll revenue to improve the distributional impacts of congestion pricing. Assuming tolls are non-negative is both realistic and makes it harder to generate a Pareto improvement.

$$- \left(t^{\max} - t^*\right) \alpha \min\{\delta, \hat{\delta}\} \begin{cases} 1 & t^* \leq t^{\max} \\ -\xi & t^* > t^{\max} \end{cases},$$

$$\bar{p}_{\text{toll}}(\alpha, \delta, t^*) = \frac{\xi}{1+\xi} \frac{1}{s^*} \left[\int_0^{\infty} \min\{\beta', \alpha\delta, \hat{\beta}\} n_\beta(\beta')\, d\beta' \right] \tag{3}$$

$$- \left(t^{\max} - t^*\right) \min\{\alpha\delta, \hat{\beta}\} \begin{cases} 1 & t^* \leq t^{\max} \\ -\xi & t^* > t^{\max} \end{cases},$$

where $\hat{\delta}$ and $\hat{\beta}$ are the marginal type to arrive during $[t_s, t_e]$, and $n_\delta(\delta)$ and $n_\beta(\beta)$ are the marginal distributions of δ and β.

To see the intuition behind these expressions for trip cost, I first write them as

$$\text{trip cost} = \frac{\xi}{1+\xi} \times \text{length rush hour} \times \text{censored mean of willingness to pay} - \text{adjustment for desired arrival time.} \tag{4}$$

Next, we can work through each term of (4). The ratio $\xi / (1+\xi)$ is uniquely determined by the ratio of the cost of being late to the cost of being early and is the fraction of agents who arrive before the peak of rush hour. If ξ is zero then it is costless to be late, as a result agents can wait to travel until there is no traffic or toll; everyone will be late and have a trip cost of zero. As ξ increases the costs of being late increases and so a larger share of agents arrive before the peak. Because agents care more about arriving on-time, travel times (or tolls) are higher and everyone's trip cost increases.

The next factor is the length of rush hour, where I define rush hour as the period of time when agents are traveling. This implies a broader definition of rush hour than in common parlance, with rush hour on a free route being the entire time travel times are higher than they would be in free flow conditions, not just when travel times are exceptionally long. Since I have normalized the mass of agents to one, the length of rush hour is $1/s$ on a free route and is $1/s^*$ on a priced route. A longer rush hour increases trip costs since it leads to more schedule delay and higher travel times or tolls.

The final factor of the first term is the most interesting; the integrals in (2) and (3) are the censored mean of an agent's willingness to pay in the currency the route requires—travel time on a free route or money on a tolled route—to reduce schedule delay. On the free route this is multiplied by the agent's value of time to convert from travel time to dollars.

The censoring occurs at the willingness to pay of the marginal agent who arrives at the same time as the agent whose trip cost we are considering. For an agent with $\delta < \hat{\delta}$ on a free route or $\beta < \hat{\beta}$ on a priced route this is his own willingness to pay. This means he does not care about the actual preferences of those with a higher willingness to pay; whether they are willing to pay a cent more or a thousand dollars more for the most desirable arrival times does not matter—either way they are willing to pay more than him for the most desirable arrival times. All that matters is how much of the desirable arrival time they use, and thus how much schedule delay the agent in question experiences. In contrast, he cares very much about the preferences of those whom he is willing to pay more than, since he will, in equilibrium, actually pay more than them to obtain his arrival time. If an agent is inframarginal, so $\delta > \hat{\delta}$ on a free route or $\beta > \hat{\beta}$ on a priced route, then the censoring occurs at the marginal willingness to pay of the marginal agent at the time they arrive.

The logic for this censoring is similar to that determining rents in the monocentric city model with heterogeneous agents.[20] Those with the lowest willingness to pay for a location near the central business district (CBD) (i.e., those with the lowest slope of the bid-rent curve) live at the edge of the city, and their total housing plus commuting costs do not depend on the exact preferences of those with a higher willingness to pay, just on the total mass of those with a higher willingness to pay. In contrast, those with the highest willingness to pay to live near the CBD (i.e., those with a steep bid-rent curve) actually live at the CBD, but the rent they pay depends on the preferences of those with a lower willingness to pay, and thus their total commuting plus rent costs depends on the preferences of the other agents. For any given individual, commuting costs are driven by how many people have a higher willingness to pay (and thus live closer to the CBD), while rent is driven by both the number and actual preferences of those with a lower willingness to pay. Likewise, in the bottleneck model, schedule delay costs are driven by how many people have a higher willingness to pay for a desirable arrival time, while travel times (or tolls) are determined by the number and actual preferences of those with a lower willingness to pay.

[20]For an example of such a model, see the excellent handbook chapter by Duranton and Puga (2015).

The final term is an adjustment for differences in desired arrival times. Those who want to arrive at the peak of rush hour pay the highest costs, while those who prefer to arrive further from the peak pay lower costs.

4.2. **Equilibrium When Value Pricing.** Solving for the equilibrium with two routes is more complicated because agents choose which route they take as well as their arrival time. In contrast to when pricing either all or none of the lanes, I must solve for equilibrium numerically. In Appendix B I show there exists a unique function, $\hat{\alpha}(\delta, t^*)$, which separates the space of agents' preference parameters into those on the free route and those on the priced route, which I approximate using Chebyshev collocation. I then derive additional results that inform my numerical approximation, show that the approximation error is small, setup the system of equations that define the equilibrium, and show that this system of equations has a unique solution.

5. Estimating the Distribution of Consumer Preferences

The theoretical analysis in Hall (2018) shows that whether value pricing can make all road users better off depends on agents' preferences. To assess whether this theoretically possibility is a real-world possibility, I now turn to estimating the distribution of agents' preferences, along with other relevant parameters. In Section 6, I use these results to evaluate the distribution and size of the welfare gains from congestion pricing.

The main structural object I estimate is the joint distribution of agents' inflexibility, value of time, and desired arrival time. My approach is to split the population into two categories using a measure of whether an agent is, broadly speaking, flexible or inflexible; then, within each category I estimate the marginal distributions of the three preference parameters: inflexibility, value of time, and desired arrival time. Having done so, I combine these marginal distributions into a joint distribution by assuming each preference parameter is independent of the others.[21] This means the correlations between preference parameters manifest themselves through differences in the marginal distributions *between* categories, rather than through the correlations *within* a category.

[21] In Appendix C.2 and C.3 I conduct two tests of the assumption of independence of marginal distributions within a category. In both cases I fail to reject the hypothesis that they are independent.

5.1. **Data.** I estimate this joint distribution for road users on a segment of California State Route 91 (SR-91). The segment I focus on is thirty-three miles long and runs from the center of Corona to the junction of SR-91 and I-605. I choose this specific segment because it roughly represents the median commute for those living in Corona who use SR-91.[22]

I use data from three sources. The first, the 1999 wave of the California Polytechnic State University's State Route 91 Impact Study (Sullivan 1999), is a surveys of road users who use SR-91. I use this data to estimate the fraction of agents who are flexible, the distribution of the value of time, and the distribution of desired arrival times.

The second data set is the 2009 National Household Travel Survey ("NHTS", U.S. Department of Transportation 2009), which I use to confirm that my estimates from the SR-91 Impact Study are similar to what I would estimate for other large metropolitan statistical areas (MSAs).[23]

The final data set is the California Department of Transportation's Performance Measurement System ("PeMS", 2014). PeMS includes road detector data from almost all the highways in California. From this data set, I calculate travel times for every business day in 2004.[24] I use these travel times to estimate the distribution of inflexibility.

5.2. **Fraction of Road Users Who Are Flexible.** The first task is to estimate the relative sizes of the two categories of agents.

The SR-91 Impact Study contains two measures of flexibility: one focuses on the driver's *typical* trip, the other on a *specific* recent trip. The first row of Table 1 shows that 57% of road users on SR-91 report that they typically leave early or late to avoid traffic congestion, and the second row shows that 43% of road users could choose what time they arrived at their destination for a specific peak period trip. The strength of the first measure is that it asks whether the road user takes an action which reveals their flexibility, while the strength of the second is that it is about a specific recent trip and so better reveals the fraction of road users on a given morning who are flexible.

[22]In Appendix C.4 and C.5 I discuss how the assumptions that everyone has the same commute, and that it is possible to price the entire commute (i.e., there are not local roads at the start or end of the trip) affect my results.

[23]I define a large MSA as one with a population above three million.

[24]I define a business day as a weekday which is not one of the ten United States federal holidays.

TABLE 1. Fraction of drivers and trips that are flexible

Fraction of ...	SR-91	Large MSAs
Drivers who typically leave early or late to avoid traffic	.57 [.55, .60]	
Trips on interstate during morning where drivers can choose arrive time	.43 [.40, .47]	.35–.60 [.32, .62]
Trips on interstate to work where drivers can choose arrival time	.50 [.47, .53]	.54 [.51, .57]

Notes: 95% confidence intervals in brackets. Confidence intervals in the second column calculated using jackknife-2 replicate weights. A trip is flexible if the driver and all passengers can choose when to arrive at their destination, unless the destination is the driver's home, in which case they must be able to choose their departure time. A trip is a series of trip segments which ends when the driver stays at one destination for more than thirty minutes.

While the NHTS does not ask the same questions as the SR-91 Impact Study, it does allow me to make some comparisons between drivers on SR-91 and those in other large MSAs. The NHTS only asks individuals if they can choose their arrival time for work trips, rather than for all trips. For the sake of making a clean comparison between drivers on SR-91 and in the rest of the United States, the third row of Table 1 reports estimates of how many road users can choose when to arrive at work from both data sets.[25] I find that the fraction of workers who are flexible on SR-91 is similar to the fraction in other large MSAs.

I can estimate the fraction of all trips during the morning that are flexible using the NHTS if I make assumptions about what kinds of non-work trips are flexible.[26] Doing so leads to the range of estimates reported in the second column of the second row. The bottom of the range comes from assuming no non-work trips are flexible, while the top comes from assuming that other trips where the driver probably has control over when it begins (such as shopping, doctor's appointments, and visiting friends) are flexible. My estimate of the fraction of

[25]I limit the NHTS sample to those who travel on the interstate, so that they are similar to those traveling on SR-91. While SR-91 is not an interstate, it is a limited access highway and so indistinguishable in all but signage from an interstate.

[26]The morning is defined in the SR-91 Impact Study as 4–10 AM and so for consistency I maintain that definition with the NHTS. I also continue to limit the NHTS sample to those who travel on the interstate.

trips on SR-91 that are flexible falls roughly in the middle of the range of estimates for large MSAs.

I use the specific-trip measure as my definition of which agents are in the flexible category. Doing so gives more conservative estimates for the maximum fraction that can be priced while generating a Pareto improvement, as well as for the private welfare gains from pricing a given fraction of the lanes. All other results are largely unaffected by which definition of flexibility I use. The results using the typical-trip measure of flexibility are reported in Appendices C and D.

5.3. **Distribution of the Value of Time.** Given this broad categorization of road users into flexible and inflexible, I now estimate the marginal distributions of value of time, desired arrival time, and inflexibility within both categories. I start by estimating the distribution of the value of time. To estimate the distribution of the value of time I first map household income into value of time and then fit a long-normal distribution to the data using maximum likelihood. I do this separately for the two broad categories for the flexible and inflexible road users.

To map household income to value of time, I use the following U.S. Department of Transportation formula: an individual's value of time is half their hourly household income, which is their annual household income divided by 2,080 hours per year (Belenky 2011, 12).[27,28] While it would be preferable to use annual individual income, or better yet, individual wages, the SR-91 Impact Study and NHTS do not contain this information.

Using this formula means I underestimate the welfare gains from congestion pricing and overstate the difficulty in obtaining a Pareto improvement. This is due to two standard results in the literature on the value of travel time. The first is that drivers particularly dislike traveling in congested traffic, valuing a reduction in it at above half their wage (Small and Verhoef 2007, 53), which means I underestimate the welfare gains. The second is that value of time does not increase proportionally with income (ibid.), and so by assuming it increases proportionally with income I overestimate the variance in value of time. This biases my results against finding a Pareto improvement.

[27]The U.S. Department of Transportation uses this formula to estimate a median value of time based on median household income, I am going further in using it by applying it to individuals.
[28]There is a large literature estimating the mean or median value of time, which generally finds it is half the mean or median wage, though it is higher when roads are congested. See Small and Verhoef (2007, 53) for a literature review.

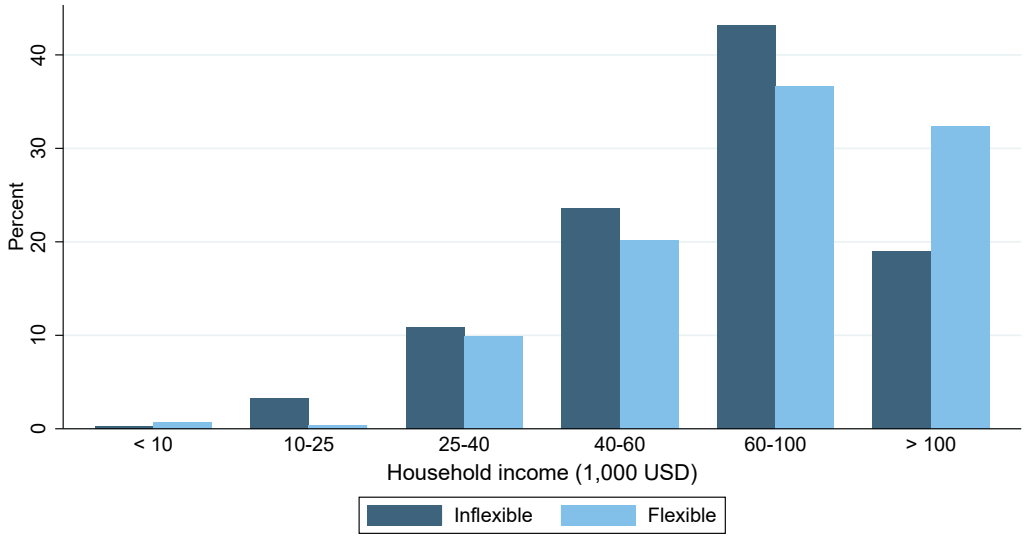

FIGURE 2. Distribution of household income by flexible/inflexible. Data from SR-91 Impact Study. Dollars are in 1998 dollars. Flexible/inflexible defined by whether the traveler typically leaves early or late to avoid traffic.

The income data is categorical. The SR-91 Impact Study reports six bins and is plotted in Figure 2. The NHTS data is more detailed, with 18 bins. The lowest 16 bins are each 5,000 dollars wide, and the top two bins are 80,000–100,000 and 100,000 or more.

I fit a log-normal distribution to the data using maximum likelihood. To write the likelihood function, define y_i as the income category for observation i, and $h(y_i)$ and $l(y_i)$ as the highest and lowest incomes within category y_i. Further define c as the function that converts household income into an estimate of value of time using the formula from Belenky (2011) and adjusts for inflation using the consumer price index, using 2012 as the base year, and F as the cumulative distribution function for a log-normal distribution which depends on the parameter vector $\vec{\theta}$. Using this notation, I write the likelihood of observing the data as

$$\mathcal{L}\left(\vec{\theta}|\vec{y}\right) = \prod_{i=1}^{N}\left[F\left(c(h(y_i))|\vec{\theta}\right) - F\left(c(l(y_i))|\vec{\theta}\right)\right]. \tag{5}$$

I estimate the parameters of the distribution of the value of time by finding the parameters $\vec{\theta}$ which maximize (5).

I estimate the distribution of the value of time separately for the flexible and inflexible categories, as mentioned earlier. While the levels of these estimates

TABLE 2. Distribution of value of time for morning highway users

	SR-91	Large MSAs
Flexible		
Median	25.95	26.05
	(0.90)	(0.34)
Interquartile range	20.0	32.21
	(1.8)	(0.89)
N	303	7,059
Inflexible		
Median	22.16	22.52
	(0.56)	(0.27)
Interquartile range	15.19	24.55
	(0.87)	(0.59)
N	433	4,270
Rank correlation†	0.200***	0.157***
	(0.055)	(0.037)

Notes: Standard errors in parentheses. Standard errors for SR-91 estimates are calculated by bootstrapping. The data for large MSAs are weighted using individual weights and their standard errors calculated using jackknife-2 replicate weights. I convert household income to value of time using a formula from the USDOT (Belenky 2011), adjust dollar amounts to 2012 dollars using the CPI, and then fit the categorical data to a log-normal distribution using maximum likelihood. For large MSAs I use the most generous definition of flexibility, which assumes certain non-work trips are flexible.

† Goodman and Kruskal's γ between income and flexibility.

*** $p < .001$

matter for valuing the time saved by congestion pricing, as Hall (2018) shows, it is the correlation between value of time and flexibility that affects our ability to obtain a Pareto improvement.

The first column of Table 2 reports the results from estimating the distribution of the value of time using the SR-91 Impact Study. Consistent with Figure 2, I find that flexible agents on average have higher values of time than the inflexible. The Goodman and Kruskal's rank correlation between flexibility and income (reported in the last row) is 0.20, which can be roughly interpreted as meaning that three

times out of five a randomly selected flexible agent has a higher value of time than a randomly selected inflexible agent.[29]

The second column in Table 2 present the results from estimating the distribution of the value of time in large MSAs using the NHTS. In this column I use the most generous form of the NHTS definition of flexibility, which assumes certain non-work trips are flexible.

As with the estimates of the fraction of agents in the flexible category, the results for SR-91 are similar to other large MSAs. Comparing both columns of Table 2 shows similar estimates for the median value of time for each category, and a relatively similar rank correlation. While my estimates of the interquartile range are much larger for large MSAs than for SR-91, this is probably because the set of people who use a particular road is likely to be more homogeneous than the set of people living in a particular MSA.

As a benchmark, I compare my estimates of the distribution of the value of time to those of Small et al. (2005), who use more detailed data and more sophisticated methods to measure the distribution of value of time for road users on SR-91. While they do not estimate the distribution separately for flexible and inflexible agents, I can compare their results to those from my relatively simple method when pooling the flexible and inflexible agents. Adjusting for inflation, they find that the median value of time is \$29.54 and the interquartile range is \$10.47, while I find a median of \$23.58 and an interquartile range of \$17.06. As expected, I underestimate the median and overestimate the interquartile range.

5.4. **Distribution of Desired Arrival Time.** The next task is to estimate the length of time over which agents desire to arrive. What I care about is the time when agents *desire* to arrive at the *highway exit*, but what I observe in the data is when they *actually* arrive at their *destination*; this means the data is informative about the shape and spread in the distribution of desired arrival times for inflexible agents, but not for the position of the distribution and not for flexible agents.

Because I observe *actual* arrival times instead of *desired* arrival times, I must focus on those whose actual arrival times are their desired arrival times—that is, those in the inflexible category, who are unable to choose their arrival time.

Because I observe arrival times at agents' *destinations* rather than at the *highway exit*, the underlying distribution I want to recover (the distribution of inflexible agents' desired arrival times at the highway exit) is a smoothed and shifted version

[29]This interpretation would be exact if there were no ties in the data.

of the distribution I observe (the distribution of inflexible agents' actual arrival times at their destination). The distribution is smoothed because the distance agents must travel from the end of the highway to their destinations varies, and so among those who want to reach their destination at 7:00 AM, there are some who want to reach the end of the highway at 6:40 AM and others who want to reach it at 6:55 AM The distribution is shifted because agents want to exit the highway earlier than they want to arrive at work. To correct for this shift, I estimate the first desired arrival time, t_s, as part of the structural estimation of the distribution of flexibility.

In Appendix C.6, I provide evidence that it is reasonable to approximate the distribution of desired arrival times with a uniform distribution, with the approximation error primarily occurring for very early and very late arrivals. I also explain why I do not expect this assumption to affect my results.

I estimate the spread in the distribution of desired arrivals times by matching the observed 10th and 90th percentiles to their expected values. I show in Appendix C.7 that this procedure gives me an unbiased estimate of the length of time over which inflexible agents wish to arrive, $t_e - t_s$. I estimate that $t_e - t_s$ is 4.40 hours, as is reported in the first row of the first column of Table 4.

As I do not observe the distribution of desired arrival times for those who are flexible, I assume it is the same as the distribution for those who are inflexible. This assumption is relatively harmless. When an agent is marginal, his desired arrival time determines whether he is early or late, but not his actual arrival time. Ascribing the wrong desired arrival time to an agent who is always marginal does not affect the equilibrium or the change in the agents' trip costs due to congestion pricing.[30] Fortunately, most agents in the flexible category are always marginal.

5.5. **Distribution of Inflexibility.** The model provides a mapping between model parameters and travel times $T(t)$. By inverting this mapping, I estimate the remaining parameters: the distribution of inflexibility, $N_\delta(\delta)$; the ratio of the cost of being early to late, ξ; the length of rush hour on a free route, $1/s$; the first desired arrival time at the highway exit, t_s; and free flow travel times, T_f.

I am only able to estimate the distribution of inflexibility for those road users who do not arrive on-time. For all other road users, I only obtain a lower bound. This follows from Lemma B.3 and is due to the fact that $T(t)$ does not reflect the preferences of the inframarginal road users.

[30] It affects the level of their trip costs, but in a consistent way so that it differences out.

Because I am unable to observe a portion of the distribution of inflexibility, I make assumptions about its shape, and then test the sensitivity of my results to these assumptions. I assume that the distribution of inflexibility for those agents who are in the flexible category, N_δ^f, is uniform on $\left[0, \tilde{\delta}\right]$; and the inflexibility of those in the inflexible category, N_δ^i, has a beta$(5, 0.5)$ distribution transformed to have support $\left[\tilde{\delta}, 1\right]$. This means I am assuming most of the inflexible agents are very inflexible, as most of the weight of N_δ^i is near one and its mode is one. The assumed form of N_δ^i does not affect the estimation of the other parameters; however, it does affect the counterfactual results, and so Appendix D reports the counterfactual results for a wide variety of assumptions about N_δ^i.[31]

I estimate the remaining parameters $\vec{\theta} = \left\{\tilde{\delta}, \xi, s, t_s, T_f\right\}$ by using the Generalized Method of Moments (GMM) to choose $\vec{\theta}$ to best match the model-predicted travel times to the empirical travel times calculated from the PeMS data set. The estimation uses my estimates of the fraction of agents who are flexible, ϕ, and the length of desired arrivals, $t_e - t_s$, from Sections 5.2 and 5.4. While I could estimate ϕ and $t_e - t_s$ as part of the GMM routine, and doing so would allow me to match $T(t)$ better, I estimate them separately because I have natural measures of them and want to match those particular "moments" exactly.

Letting x_{ij} be the observed travel time on SR-91W from the center of Corona to the junction of SR-91 and I-605 for arrival time i on day j, I write the moment conditions as

$$T\left(t_i; \tilde{\delta}, \xi, s, t_s, \phi, t_e - t_s\right) + T_f = \sum_j x_{ij}/250 \quad \text{for } t_i \in \{4{:}00,\ 4{:}05, \ldots,\ 10{:}00\}\ ,$$

where $T\left(t_i; \tilde{\delta}, \xi, s, t_s, \phi, t_e - t_s\right)$ is defined by (B.19) in Appendix B.

While each parameter in $\vec{\theta}$ is chosen to best match the model's predicted $T(t)$ to the entire time series of empirical travel times, each parameter also directly maps into a particular feature of the predicted time series of travel times, and thus the estimate for each parameter is strongly affected by a particular feature of the empirical time series of travel times. These relationships are reported in Table 3.

To test the restrictiveness of my functional form assumptions, I fit a relaxed version of the model to the data nonparametrically. After relaxing the functional form assumptions for the distributions of inflexibility and desired arrival times, as

[31] I plot whether pricing generates a Pareto improvement for all combinations of the parameters of the beta distribution in $[.1, 10] \times [.1, 10]$, and report the largest welfare loss, and average annual social and private welfare gains, for six representative sets of parameter values.

TABLE 3. Which features of the data identify which parameters

Parameter	Notation	Feature of data which identifies parameter
Distribution of inflexibility	$n_\delta(\delta)$	Distribution of the slope of $T(t)$
Ratio of schedule delay costs late-to-early	ξ	Ratio of the average slope after the peak to the average slope before the peak
Free flow travel time	T_f	Average travel time between 4 a.m. and when travel times start climbing
Length of rush hour	$1/s$	Length of time when travel times are above T_f
First desired arrival time	t_s	Time when the slope of $T(t)$ stops changing because the marginal type becomes constant at t_s when desired arrival times are uniformly distributed

well as the assumption that the ratio of the cost of being late to early is the same for all agents, the theory still imposes three sets of constraints on travel times. Letting $i^{\max}$ index the start of the five-minute period in which the peak of rush hour occurs, the constraints are as follows:

(1) Travel times are positive:

$$T_i > 0 \ \forall i.$$

(2) Travel times are increasing before the peak and decreasing after:

$$\begin{aligned} T_i &\geq T_{i-1} \quad \forall i \leq i^{\max} \quad \text{and} \\ T_i &\leq T_{i-1} \quad \forall i > i^{\max}. \end{aligned}$$

(3) Travel times are convex before the peak and convex after the peak:

$$\frac{T_i - T_{i-1}}{t_i - t_{i-1}} \geq \frac{T_{i-1} - T_{i-2}}{t_{i-1} - t_{i-2}} \qquad \forall i \notin \{i^{\max}+1, i^{\max}+2, i^{\max}+3\}.$$

The first constraint is never binding and the third constraint makes the second constraint redundant for all but the first and last arrival times.

To fit the relaxed model to the data nonparametrically, I find the travel times, T_i, as well as the index of the five-minute window in which the peak of rush hour

TABLE 4. Remaining parameter estimates

	GMM	Nonparametric
Length of desired arrivals (hours) $(t_e - t_s)$	4.40	4.33
	(0.22)	(0.22)
First desired arrival time (hours) (t_s)	5.556	5.333
	(0.063)	(0.068)
Length of rush hour on free route (hours) $(1/s)$	7.74	8.00
	(0.40)	(0.33)
Maximum inflexibility of flexible agents $(\bar{\delta})$	0.228	—
	(0.042)	
Ratio of schedule delay costs late to early (ξ)	0.411	0.403
	(0.033)	(0.039)
Free flow travel time (minutes) (T_f)	36.71	36.69
	(0.90)	(0.77)

Note: Bootstrapped standard errors in parentheses. The estimate in the first row of the first column comes from fitting the largest and smallest observations of the trimmed sample of the inflexible agents' desired arrival times to the expected value of their order statistics (N = 488). The last five rows of the first column report the GMM estimates (N = 250). The second column reports nonparametric estimates, which come from finding the predicted travel times that best meet a minimum set of restrictions implied by the model, and then estimating parameters from these predicted travel times (N = 250). The nonparametric estimates of t_s and $t_e - t_s$ require assuming desired arrival times are uniformly distributed.

occurs, $i^{\max}$, which minimize the GMM criterion subject to the three sets of constraints above. I then use the predicted $T(t)$ from the nonparametric estimation and the relationships from Table 3 to nonparametrically estimate $t_s, t_e - t_s, 1/s, \xi$, and T_f. I am unable to estimate the distribution of inflexibility nonparametrically.[32]

Table 4 reports the GMM and nonparametric estimates, which are very similar. In particular, the nonparametric estimate of the length of desired arrivals is almost the same as the estimate from Section 5.4, even though it is estimated from travel times rather than survey data.

The *length of rush hour* is the period of time when travel times are higher than they would be in free flow conditions, not just when they are exceptionally long.

[32]See Appendix C.8 for details on the nonparametric estimation and Appendix C.9 for a discussion of why I cannot estimate N_δ nonparametrically.

Using this definition, I estimate that rush hour is more than seven and a half hours long, starting before five in the morning and not ending until a little after noon.

I estimate that the inflexibility of those in the flexible category is uniformly distributed on $[0, 0.228]$ and that the ratio of the cost of being late to the cost of being early is 0.4. This last result means the cost of being late is *less* than the cost of being early; while this appears unreasonable, it is largely a result of how I estimate this ratio. In a model where travelers differed in their relative cost of arriving early vs. late, those for whom arriving late is relatively inexpensive would arrive late, while those for whom arriving early is relatively inexpensive would arrive early. If this is the case, then these estimates are best interpreted as saying that the marginal driver who is late pays lower schedule delay costs than the marginal driver who is early; there is nothing unreasonable about this. As discussed in footnote 15, allowing travelers to have different relative costs of arriving early vs. late would only affect my results if there are people who switch from arriving early to arriving late, or vice-versa, when tolls are added to the road.

Furthermore, being late does not necessarily mean literally arriving late to an appointment, but can mean a traveler prefers to go to the doctor at 9 AM but instead schedules the appointment for 11 AM to avoid traffic. He arrives exactly on-time to his 11 AM appointment, but still has schedule delay costs. While being late to an appointment is likely more costly than being early, scheduling an appointment later than desired may be less costly than scheduling it early. My estimates are of these long-run schedule delay costs.

The empirical travel times along with the predicted travel times from both methods are shown in Figure 3. Both predicted sets of travel times match the data well and are difficult to tell apart. Making the additional assumptions about functional forms only increases the root GMM criterion by 7.7%. The small difference in the root GMM criterion of the nonparametric and GMM estimates, as well as the similarity in their parameter estimates, suggests that it is innocuous to make these additional assumptions.

6. Counterfactuals

Given the estimated distribution of driver preferences from Section 5, I use the results of Section 4 to solve for the equilibrium under counterfactual congestion pricing regimes. This allows me to estimate the distributional and aggregate

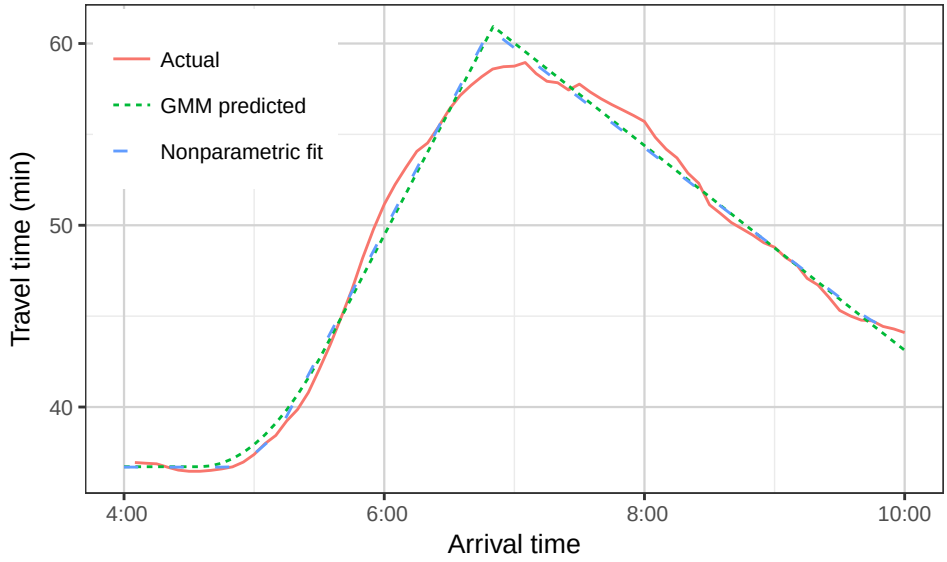

FIGURE 3. Actual versus predicted travel times. Data from PeMS.

welfare effects of pricing a portion or all of the lanes. I conclude this section with a variety of sensitivity checks on these results.

The final parameter needed to evaluate counterfactuals is the amount throughput falls once a queue forms. Defining $s = (1 - \theta)s^*$, this parameter is θ, and I call θ the *size of the throughput drop*. The traffic engineering literature estimates that queues at bottlenecks reduce throughput by roughly 10%, while queue spillovers reduce throughput by 25%.[33] I report results for both of these levels of throughput drop, as well as the midpoint between them. For the sake of the discussion, I focus on the middle case. The results for all three sizes of the throughput drop follow the same patterns, though the specific numbers are different.

6.1. **Distributional and Aggregate Welfare Effects.** Table 5 reports the largest annual welfare loss, the average annual welfare effects, and the decomposition of the welfare effect of pricing all or part of the lanes of the highway. The headline result is that pricing a portion of the lanes generates a Pareto improvement, while pricing all of them significantly hurts some agents.

While pricing all the lanes raises significant revenue, it is not enough for a uniform rebate to make pricing the entire road yield a Pareto improvement. The worst-off agent is hurt by $2,390 per year, which is 70% larger than annual toll revenue per capita. This means using the revenue to make pricing the entire road

[33] See Hall (2018) for a summary of this literature.

TABLE 5. Average annual welfare effects of congestion pricing

Size of throughput drop (%)	10		17.5		25	
Fraction of lanes priced	1	0.25	1	0.5	1	0.5
Largest welfare loss ($)†	3420	0	2390	0	1590	0.0
	(420)	(120)	(320)	(55)	(290)	(0.0021)
	[0.0]	[0.74]	[0.0]	[0.94]	[0.01]	[1.0]
Welfare gains ($)						
Social	2270	1010	2400	1740	2510	1910
	(280)	(140)	(290)	(230)	(290)	(250)
Private	490	310	1080	760	1580	1020
	(240)	(85)	(240)	(150)	(290)	(200)
Reduction in travel time (hours)	76.5	18.1	76.5	40.5	76.5	45.9
	(9.1)	(3.5)	(9.1)	(6.0)	(9.1)	(7.0)
Reduction in travel time costs ($)	1960	820	1960	1390	1960	1490
	(250)	(130)	(250)	(200)	(250)	(210)
Reduction in schedule delay (hours)	113.6	30.7	199	108.9	284	162
	(7.3)	(2.0)	(13)	(7.0)	(18)	(10)
Reduction in schedule delay costs ($)	305	195	438	342	543	417
	(41)	(28)	(52)	(42)	(63)	(50)
Tolls Paid ($)	1780	700	1320	970	930	888
	(200)	(82)	(160)	(110)	(150)	(98)

Notes: Bootstrapped standard errors in parentheses. The fraction of bootstrapping iterations for which pricing a given fraction of the road yields a Pareto improvement is in brackets. I assume two trips per working day and 250 working days per year. Social welfare gains are the sum of the reduction in travel time costs and the reduction in schedule delay costs; and they do not include the value of saving gasoline or reducing pollution. Private welfare gains are social welfare gains minus the private cost of the tolls paid. Numbers in the table do not add up exactly due to rounding.

† The largest welfare loss is not an average, but the maximum annual welfare loss.

generate a Pareto improvement requires spending it in a way that targets those who are harmed. This is difficult to do, as Foster (1975) notes.

Value pricing generates a Pareto improvement even before using the revenue; however, doing so requires giving up some of the potential social welfare gains.

Not pricing all the lanes leaves some lanes congested and with lower throughput; reducing the social welfare gain by 30% relative to pricing all the lanes. That said, if by making congestion pricing yield a Pareto improvement we can actually implement congestion pricing then we are trading $660 per person per year of potential, unrealized, welfare gains for $1,740 per person per year of actual welfare gains.

Fortunately, the welfare cost of not pricing all the lanes is relatively small, as pricing half the lanes yields more than half the potential welfare gains. While the travel time savings from tolling are essentially proportional to the share of the lanes priced, the *value* of the travel time savings is more-than-proportional because those with a high value of time are choosing to travel on the priced lanes. Thus pricing half the lanes saves the most valuable half of the travel time and captures over 70% of the value of the reduction in travel time. The same pattern repeats itself with the reduction in schedule delay. This allows us to capture a more-than-proportional share of the available social welfare gains when pricing a portion of the lanes.

When deciding what fraction of the lanes to price we are trading off efficiency and distributional concerns. Figure 4 shows this trade off when θ, the amount throughput drops once a queue forms, is only 10%. The largest drop in the maximum harm comes from leaving at least some of the lanes unpriced, because the inflexible poor prefer a congested but free option over needing to pay a toll. If we are willing to relax the requirement that pricing must generate a Pareto improvement and instead put some bound on the maximum harm done, then we can reap a greater portion of the potential welfare gains.

The welfare gains available from congestion pricing are large; even in the conservative case they are over $1,000 per agent per year. In the middle case pricing half of the lanes would be equivalent to increasing the median income of these agents by over 3.5%, and pricing all the lanes would increase median income by over 5%. Most of the welfare gain comes from changing the currency used to pay for desired arrival times from time to money. The time spent in traffic is a social loss while the money spent on tolls is just a transfer. Most of this portion of the welfare gains accrues to whomever gets to keep the toll revenue. However, a significant amount of the welfare gains goes to the agents themselves. Even if the toll revenue is wasted the average agent will be $760 better off each year due to value pricing.

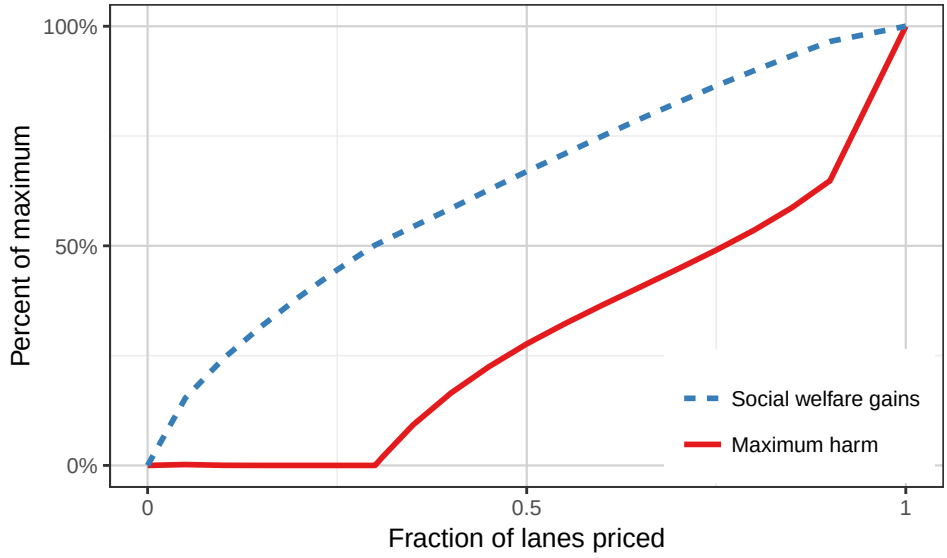

FIGURE 4. Trade-off between maximum harm and social welfare gains when throughput drops 10% once a queue forms.

The reduction in schedule delay, while large in magnitude, accounts for small share of the social welfare gains. It accounts for a small share since those who are very inflexible are already arriving on-time, and so the reductions in schedule delay occur primarily for those who have flexible schedules, and thus do not value the reduction highly. In contrast (and as discussed earlier), when value pricing the travel time savings are concentrated among those with the highest value of time.

Figure 5 shows who is helped or hurt due to pricing all (Panel A) or half (Panel B) of the lanes.[34] The agents harmed the most by pricing all the lanes are the inflexible poor (in the bottom right of Panel A)—those who need to arrive to work exactly on-time and who would strongly prefer to pay with their time to do so instead of their money. The curve of darkest red in the lower right of Panel A lies along the curve $\alpha = \hat{\beta} \cdot \delta$; these are the agents who were able to arrive exactly on time when the road was free, but when the road is priced they are displaced by flexible rich agents who start arriving during the peak. The inflexible rich (in the upper right) are the best off; when the road is free they arrive on-time but bear large travel time costs, and they are delighted to pay with their money instead of their time. The flexible (on the left) are not very affected by adding tolls; they avoided paying with travel time by arriving off-peak and they will avoid paying

[34]Figure 5 does not show the 8% of agents with a value of time above fifty dollars an hour.

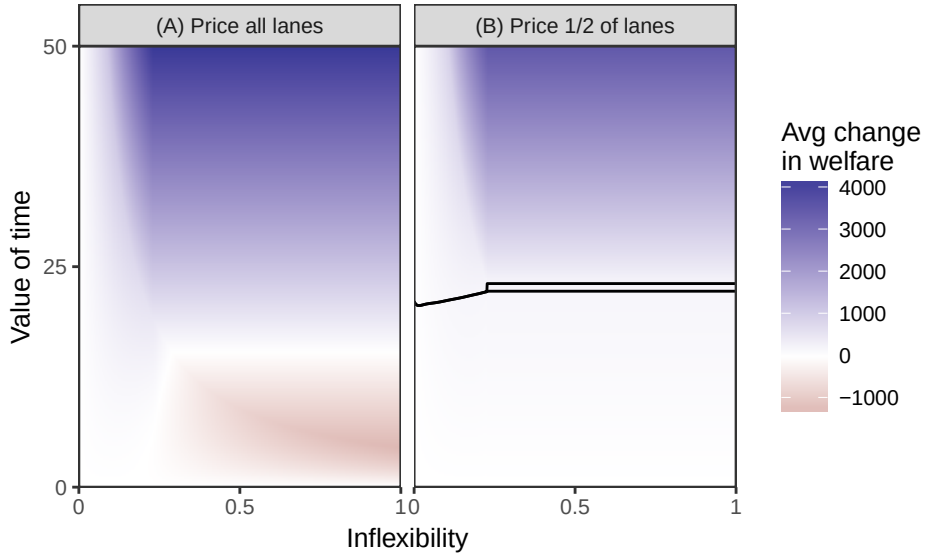

FIGURE 5. Annual change in welfare, averaged by type, when throughput drops 17.5% after a queue forms. The black lines in Panel B are the maximum and minimum values of $\hat{\alpha}(\delta, t^*)$ for each δ.

with money by continuing to arrive off-peak. They are better off since they have a little less schedule delay, but as they are flexible they do not value the reduction highly.

Pricing just half of the lanes preserves the ability of the poor to pay with their time, and so, as Panel B shows, avoid hurting the inflexible poor. Doing so reduces the benefits to the inflexible rich, but generates a Pareto improvement.

Recall that, given the strong correlation between value of time and income, I use "rich" as a shorthand for "those with high values of time" and "poor" as shorthand for "those with low values of time". In reality, there are some poor drivers with high values of time and some rich drivers with low values of time. This means that there are some inflexible poor drivers who receive large welfare gains from pricing.[35]

Panel B also shows which agents are on which route. The black lines are the maximum and minimum values of $\hat{\alpha}(\delta, t^*)$ for each δ, and so separate the space of types into those on the priced route and those on the free route. Those above both lines are on the priced route, those below both are on the free route, and those types between the two lines have members on both routes.

[35]Empirically, 15% of those with household incomes less than $40,000 report using the priced route on CA SR-91, while 44% of those with household incomes above $100,000 report doing so.

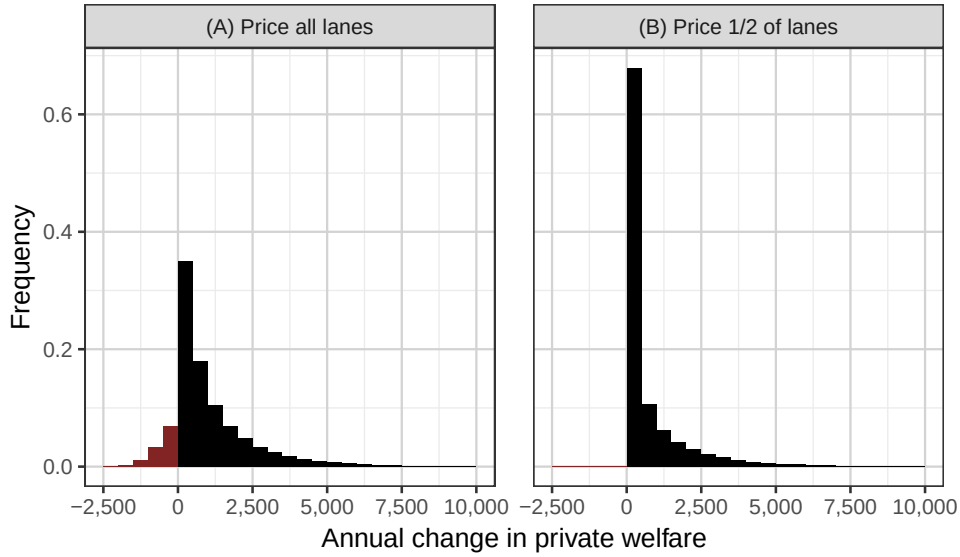

FIGURE 6. Distribution of annual changes in private welfare when the throughput drop is 17.5%.

In both panels of Figure 5 the change in trip cost is constant for a given value of time across a range of high levels of inflexibility. This occurs for the same reason $\hat{\alpha}(\delta, t^*)$ is constant when $\delta > \hat{\delta}$ and $\hat{\alpha}(\delta, t^*) > \hat{\beta}/\hat{\delta}$: if an agent is inframarginal regardless of whether the road is free or priced, then he always arrives exactly on time. When an agent arrives on time, he does not experience any schedule delay costs, and so his actual inflexibility does not affect his trip cost or the change in his trip cost. Thus two agents with the same value of time and with sufficiently high, though different, levels of inflexibility have the same change in their trip cost.

Figure 6 plots the distribution of annual changes in private welfare. It shows that when pricing all the lanes a significant group of road users are worse off, but that when pricing a portion of the lanes everyone is better off. It also shows that not hurting anyone comes at the cost of reducing benefits to others.

6.2. **Effect on Travel Times and Tolls.** Table 6 reports peak and average excess travel times (i.e., the additional travel time due to congestion) and tolls in a variety of counterfactuals. These are averaged over agents rather than time, which is why the table shows that average travel times are higher when pricing a portion of the lanes.[36] Value pricing increases the typical agent's travels times, but also reduces

[36]The difference in average excess travel time when weighting by arrival time or agent occurs because there are now times on the free route when travel times are zero, but as no one travels at these times they are not included in the average travel time experienced by agents.

TABLE 6. Travel times and tolls

Size of throughput drop (%)		10		17.5		25	
Fraction of lanes priced	0	1	0.25	1	0.5	1	0.5
Excess travel times (min)							
Average	9.2	0	9.6	0	9.6	0	8.58
	(1.1)	(0)	(1.1)	(0)	(1.0)	(0)	(0.93)
Peak	23.3	0	23.1	0	22.2	0	19.7
	(2.6)	(0)	(2.5)	(0)	(2.4)	(0)	(2.2)
Toll ($)							
Average	0	3.56	5.18	2.64	3.55	1.86	3.11
	(0)	(0.41)	(0.60)	(0.32)	(0.39)	(0.31)	(0.34)
Peak	0	9.2	12.6	6.48	8.55	4.36	7.32
	(0)	(1.2)	(1.4)	(0.87)	(0.93)	(0.77)	(0.85)

Notes: Bootstrapped standard errors in parentheses. Averages are calculated over agents, not over time.

their schedule delay, and so they are better off. Comparing travel times for any given arrival time shows that value pricing reduces travel times at every point in time, on average by between 1.3 and 22%, depending on what size of throughput drop (θ) I use. This confirms our finding that value pricing generates a Pareto improvement, as everyone in the free lanes has lower travel times, so is better off, and those in the priced lanes could have stayed in the free lanes but chose not to, so must be better off as well.

Tolls are higher when pricing only some of the lanes. This occurs because when there are fewer agents in the priced lanes, the marginal agent has a higher value of time. The tolls reflect the marginal agents' preferences and so are higher. This is also why tolls paid (i.e., annual per capita toll revenue) in Table 5 are more than proportional to the fraction of lanes priced.

6.3. **Extrapolating to the Rest of the United States.** To see the possible welfare gains from implementing congestion pricing across the United States, I extrapolate my estimates of the private and social welfare gains, both per road user and in total, from congestion pricing to all the metropolitan statistical areas (MSAs) in the United States. To do so, I assume that the distribution of traveler preferences is

the same across cities, but allow the severity of congestion to differ. For each city, I solve for the ratio of the number of travelers to road capacity that generates the estimated measures of the severity of congestion in Schrank et al. (2012). Solving for equilibrium when the road is free, and when all or part of the lanes are priced, allows me to calculate the private and social welfare gains per person-mile of travel. I then use data from U.S. Department of Transportation (2009) and estimates from Margiotta et al. (1994) and Schrank et al. (2012) to estimate the number of person-miles exposed to congestion (both total and for the typical trip). The product of these gives the total social and private welfare gains from pricing, both total and for the typical trip. The full details are in Appendix D.[37]

Estimates of the private and social welfare gains from congestion pricing for all the MSAs in the United States are reported in Appendix Tables D.1 and D.2. I find that pricing half the lanes on all urban highways would increase social welfare by over $30 billion per year. By way of comparison, recent estimates of the cost of congestion have included $37.5 billion (Winston and Langer 2006), $82 billion (Couture et al. 2018), and $160 billion (Schrank et al. 2015); while these estimates are larger than mine, they are measuring the cost of congestion on all the lanes of all roads for all hours.

The social welfare gains for the typical commuter average $850 per year, with slightly more than half the welfare gains accruing directly to the commuters. The social welfare gains are smaller for the typical urban road user than for those on SR-91 because SR-91 is among the most congested highways in America and those who use it have longer-than-average commutes.

6.4. **Sensitivity Checks.** I conduct a variety of sensitivity checks in Appendix D. Table D.3 recreates Table 5 using the typical-trip measure of flexibility. The results are largely unchanged: value pricing generates a Pareto improvement while pricing all the lanes does not, the largest welfare loss from pricing all the lanes is 30% lower, and social welfare gains are similar. The composition of the social welfare gains does differ: tolls revenue are lower and private welfare gains are larger.

Table D.5 and Figure D.1 compare results when using different assumptions on the distribution of the inflexibility of those agents in the inflexible category.

[37] A caveat on these results is that while my assumption that local roads are a small part of the entire commute (for commuters who use the highway) fits Los Angeles well due to its dense network of highways, it may not apply as well to other cities with fewer highways.

The estimates of the size of the social and private welfare gains are unchanged, but in the most conservative case (specific-trip measure of flexibility with a small throughput drop after a queue forms) whether value pricing generates a Pareto improvement depends on the distribution assumed.

7. Conclusion

This paper has investigated the practical relevance of the theoretical possibility that a carefully designed toll applied to a portion of the lanes of a highway can generate a Pareto improvement, even before the toll revenue is spent. To do so, I extended the model of Hall (2018) to allow road users' preferences over value of time and schedule inflexibility to be continuously distributed. After solving for the equilibrium, I estimated the joint distribution of road user preferences over three dimensions: value of time, schedule inflexibility, and desired arrival time. This is the first time the distribution of schedule inflexibility has been estimated, despite its importance in structural models of congestion. I then used these estimates to evaluate the aggregate and distributional effects of congestion pricing. I found that pricing all the lanes hurts many road users, in particular the inflexible poor, but pricing half the lanes generates a Pareto improvement and increases social welfare by over $1,700 per road user per year.

It is worth highlighting that the toll that generates a Pareto improvement has four important properties. First, it must be time-varying so that it induces travelers to change when they depart. Second, it should be set to maximize, or at least increase, throughput. Notably, this is almost certainly different from the toll which maximizes profits, and likely different from the toll which maximizes social welfare. Third, it must be collected electronically, as envisioned by Vickrey (1963) and seen today in the form of E-ZPass in the Northeastern United States and similar systems elsewhere, so that collecting the toll does not reduce throughput. Fourth, unless adding tolls greatly increases throughput (i.e. more than 25%), only a portion of the lanes should be tolled.

Tolls that do not have these four properties are less likely to generate a Pareto improvement. For example, a toll that is not time-varying will either be too low at the peak, so that queues form and throughput falls, or too high off-peak, discouraging travel even when the road is uncongested. Most likely it would be both.

Fortunately, designing tolls with the first three of these properties is straightforward. The technology Vickrey (1963) envisioned for the electronic collection of tolls is in use today and since highway throughput is observable, writing pricing algorithms to maximize throughput is relatively straightforward.

Determining the share of the lanes to be priced is more challenging, and depends on the distribution of preferences and how much tolling increases throughput. As Figure 4 shows, even if the share of lanes being priced is too high, so that tolling doesn't help everyone, pricing a portion of the lanes still greatly improves the distributional effects.

There are at least four ways to further improve the distributional effects of congestion pricing. First, we can use the toll revenue, either returning it to road users by reducing other vehicle taxes or by expanding the road, or spreading the benefit more broadly by cutting sales or income taxes.[38] Second, we can include carpooling or public transportation. Both of these help by increasing throughput (as measured in number of people per unit time) while also allowing travelers to reduce the financial cost of arriving at their desired arrival using the tolled route.[39] While both public transit and carpooling take additional time, the inflexible poor (those most hurt by congestion pricing) are willing to pay in time to save money, and so both public transit and carpooling help the people congestion pricing hurts most. Furthermore, these time costs are offset (and perhaps exceeded) by the time savings from using the priced lanes. Third, we can recognize that even if some drivers are worse off on some days, they benefit from the ability to take the faster priced lanes on days they are in a hurry (i.e., agents face shocks to their preferences). Fourth, we can include the environmental benefits from reducing traffic congestion with its resulting pollution.

The potential welfare gains from value pricing are large and obtainable. Extrapolating my results to the rest of the United States suggests that pricing half the lanes on urban highways would increase social welfare by over $30 billion per year, without hurting any road users.

Appendix: Supplementary Materials

The online appendix, replication code, and data for this paper are available at `https://doi.org/10.5683/SP2/RZS1FL`.

[38] See Small (1992) for a carefully designed proposal for how to use the toll revenue.

[39] Carpooling reduces the financial cost of using the priced lanes since the toll is shared among all passengers. An explicit carpool discount can further reduce the cost of carpooling.

References

AASHTO (2005) *A Policy on Design Standards-Interstate System*, Washington D.C.: American Association of State Highway and Transportation Officials, http://dx.doi.org/10.4135/9781483346526.n55.

Abrantes, Pedro A.L. and Mark R. Wardman (2011) "Meta-Analysis of UK Values of Travel Time: An Update," *Transportation Research Part A: Policy and Practice*, 45 (1), 1–17, 10.1016/j.tra.2010.08.003.

Arnott, Richard, André de Palma, and Robin Lindsey (1993) "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 83 (1), 161–179, https://www.jstor.org/stable/2117502.

_____ (1994) "The Welfare Effects of Congestion Tolls with Heterogeneous Commuters," *Journal of Transport Economics and Policy*, 28 (2), 139–161, http://www.jstor.org/stable/20053032.

Belenky, Peter (2011) "Revised Departmental Guidance on Valuation of Travel Time in Economic Analysis,", U.S. Department of Transportation, Washington, D.C., http://www.dot.gov/sites/dot.dev/files/docs/vot_guidance_092811c.pdf.

Bloomberg Philanthropies (2018) *2018 American Mayors Survey*, New York, NY, https://www.bbhub.io/dotorg/sites/2/2018/04/American-Mayors-Survey.pdf.

California Department of Transportation (2014) *Performance Measurement System*, Sacramento, California, http://pems.dot.ca.gov.

Cassidy, Michael J., Kitae Jang, and Carlos F. Daganzo (2010) "The Smoothing Effect of Carpool Lanes on Freeway Bottlenecks," *Transportation Research Part A: Policy and Practice*, 44 (2), 65–75, 10.1016/j.tra.2009.11.002.

Cohen, Yuval (1987) "Commuter Welfare under Peak-Period Congestion Tolls: Who Gains and Who Loses," *International Journal of Transport Economics*, 14 (3), 239–266, http://www.jstor.org/stable/42748188.

Couture, Victor, Gilles Duranton, and Matthew Turner (2018) "Speed," *Review of Economics and Statistics*, forthcoming.

Currie, Janet and Reed Walker (2011) "Traffic Congestion and Infant Health: Evidence from E-ZPass," *American Economic Journal: Applied Economics*, 3 (1), 65–90, 10.1257/app.3.1.65.

Daganzo, Carlos (1996) "The Nature of Freeway Gridlock and How to Prevent It," in *Traffic and Transportation Theory*, 629–646, Lyon, France: Pergamon.

De Meza, David and J. R. Gould (1987) "Free Access versus Private Property in a Resource: Income Distributions Compared," *Journal of Political Economy*, 95 (6), 1317–1325, 10.1086/261518.

de Palma, André, Robin Lindsey, and Guillaume Monchambert (2017) "The Economics of Crowding in Rail Transit," *Journal of Urban Economics*, 101, 106–122, 10.1016/j.jue.2017.06.003.

Duranton, Gilles and Diego Puga (2015) "Urban Land Use," in Duranton, Gilles, J. Vernon Henderson, and William C. Strange eds. *Handbook of Regional and Urban Economics*, 5 of Handbook of Regional and Urban Economics, 467–560: Elsevier, 10.1016/B978-0-444-59517-1.00008-8.

Duranton, Gilles and Matthew A. Turner (2011) "The Fundamental Law of Road Congestion: Evidence from US Cities," *American Economic Review*, 101 (6), 2616–2652, 10.1257/aer.101.6.2616.

Evans, Andrew W. (1992) "Road Congestion Pricing: When Is It a Good Policy?" *Journal of Transport Economics and Policy*, 26 (3), 213–243, http://www.jstor.org/stable/20052985.

Fosgerau, Mogens and André de Palma (2013) "The Dynamics of Urban Traffic Congestion and the Price of Parking," *Journal of Public Economics*, 105, 106–115, 10.1016/j.jpubeco.2013.06.008.

Foster, C. D. (1975) "A Note on the Distributional Effects of Road Pricing," *Journal of Transport Economics and Policy*, 9 (2), 186–187, http://www.jstor.org/stable/20052404.

Giuliano, Genevieve (1992) "An Assessment of the Political Acceptability of Congestion Pricing," *Transportation*, 19 (4), 335–358, 10.1007/BF01098638.

Hall, Jonathan D. (2018) "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways," *Journal of Public Economics*, 158, 113–125, 10.1016/j.jpubeco.2018.01.003.

_____ (2019) "Improving Structural Models of Congestion," *SSRN Working Paper*, 10.2139/ssrn.3424097.

Halvorson, Randy and Kenneth R. Buckeye (2006) "High-Occupancy Toll Lane Innovations: I-394 MnPASS," *Public Works Management & Policy*, 10 (3), 242–255, 10.1177/1087724X06288331.

Johnson, M. Bruce (1964) "On the Economics of Road Congestion," *Econometrica*, 32 (1/2), 137–150, 10.2307/1913739.

Kahneman, Daniel, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone (2004) "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method," *Science*, 306 (5702), 1776–1780, 10.1126/science.1103572.

Langan, Trevor, Christiana K. McFarland, and Brooks Rainwater (2017) "State of the Cities 2017,", National League of Cities, Washington, D.C., http://www.nlc.org/SOTC.

Larsen, Bradley and Anthony Lee Zhang (2018) "A Mechanism Design Approach to Identification and Estimation," NBER Working Paper 24837, National Bureau of Economic Research, 10.3386/w24837.

Lave, Charles (1994) "The Demand Curve Under Road Pricing and the Problem of Political Feasibility," *Transportation Research Part A: Policy and Practice*, 28 (2), 83–91, 10.1016/0965-8564(94)90030-2.

Light, Thomas (2009) "Optimal Highway Design and User Welfare Under Value Pricing," *Journal of Urban Economics*, 66 (2), 116–124, 10.1016/j.jue.2009.05.003.

Lindsey, Robin and Erik Verhoef (2008) "Congestion Modeling," in Hensher, David and Kennth Button eds. *Handbook of Transportation Modelling*, 2nd edition, 417–441, New York: Elsevier, https://doi.org/10.1108/9780857245670-021.

Margiotta, Richard, Harry Cohen, Robert Morris, Jeffrey Trombly, and Andrew Dixson (1994) "Roadway Usage Patterns: Urban Case Studies,", Volpe National Transportation Systems Center and Federal Highway Administration, http://ntl.bts.gov/lib/51000/51600/51637/Roadway_Usage_Patterns_1994.pdf.

May, Anthony D, Simon P. Shepherd, and John J. Bates (2000) "Supply Curves for Urban Road Networks," *Journal of Transport Economics and Policy*, 34 (3), 261–290, http://www.jstor.org/stable/20053846.

Menendez, Monica and Carlos F. Daganzo (2007) "Effects of HOV Lanes on Freeway Bottlenecks," *Transportation Research Part B: Methodological*, 41 (8), 809–822, 10.1016/j.trb.2007.03.001.

Newell, Gordon F. (1988) "Traffic Flow for the Morning Commute.," *Transportation Science*, 22 (1), 47, 10.1287/trsc.22.1.47.

Perez, Benjamin G. and Gian-Claudia Sciara (2003) "A Guide for HOT Lane Development," FHWA-OP-03-009, Federal Highway Administration, Washington, D.C., https://ntl.bts.gov/lib/jpodocs/repts_te/13668.html.

Pigou, Arthur Cecil (1920) *The Economics of Welfare*, London: Macmillan and co., ltd., 1st edition.

Santos, Georgina and Laurent Rojey (2004) "Distributional Impacts of Road Pricing: The Truth behind the Myth," *Transportation*, 31 (1), 21–42, 10.1023/B:PORT.0000007234.98158.6b.

Schrank, David, Bill Eisele, and Tim Lomax (2012) "2012 Urban Mobility Report,", Texas A&M Transportation Institute, College Station, Texas, http://mobility.tamu.edu/ums/report/.

Schrank, David, Bill Eisele, Tim Lomax, and Jim Bak (2015) "2015 Urban Mobility Scorecard,", Texas A&M Transportation Institute, College Station, Texas, https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-scorecard-2015.pdf.

Schweitzer, Lisa and Brian Taylor (2008) "Just Pricing: The Distributional Effects of Congestion Pricing and Sales Taxes," *Transportation*, 35 (6), 797–812, 10.1007/s11116-008-9165-9.

Silva, Hugo E., Erik T. Verhoef, and Vincent A. C. van den Berg (2014) "Airlines' Strategic Interactions and Airport Pricing in a Dynamic Bottleneck Model of Congestion," *Journal of Urban Economics*, 80, 13–27, 10.1016/j.jue.2013.08.002.

Small, Kenneth A. (1982) "The Scheduling of Consumer Activities: Work Trips," *American Economic Review*, 72 (3), 467–479, http://www.jstor.org/stable/1831545.

_____ (1983) "The Incidence of Congestion Tolls on Urban Highways," *Journal of Urban Economics*, 13 (1), 90–111.

_____ (1992) "Using the Revenues From Congestion Pricing," *Transportation*, 19 (4), 359–381, 10.1007/BF01098639.

Small, Kenneth A. and Xuehao Chu (2003) "Hypercongestion," *Journal of Transport Economics and Policy*, 37 (3), 319–352, http://www.jstor.org/stable/20053940.

Small, Kenneth A. and Erik T. Verhoef (2007) *The Economics of Urban Transportation*, New York: Routledge.

Small, Kenneth A., Clifford Winston, and Jia Yan (2005) "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability," *Econometrica*, 73 (4), 1367–1382, 10.1111/j.1468-0262.2005.00619.x.

_____ (2006) "Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design," *Brookings-Wharton Papers on Urban Affairs*, 53–96, 10.1353/urb.2006.0027.

Sorensen, Paul A. and Brian D. Taylor (2005) "Review and Synthesis of Road-Use Metering and Charging Systems,", Transportation Research Board, Washington,

D.C.

Starkie, David (1986) "Efficient and Politic Congestion Tolls," *Transportation Research Part A: General*, 20 (2), 169–173, 10.1016/0191-2607(86)90044-0.

Stiglitz, Joseph (1998) "Distinguished Lecture on Economics in Government: The Private Uses of Public Interests: Incentives and Institutions," *Journal of Economic Perspectives*, 12 (2), 3–22, 10.1257/jep.12.2.3.

Sullivan, Edward (1999) *State Route 91 Impact Study Datasets*, San Luis Obispo, California: California Polytechnic State University, http://ceenve3.civeng.calpoly.edu/sullivan/sr91/.

Sullivan, Edward and Mark Burris (2006) "Benefit-Cost Analysis of Variable Pricing Projects: SR-91 Express Lanes," *Journal of Transportation Engineering*, 132 (3), 191–198, 10.1061/(ASCE)0733-947X(2006)132:3(191).

Takayama, Yuki and Masao Kuwahara (2017) "Bottleneck Congestion and Residential Location of Heterogeneous Commuters," *Journal of Urban Economics*, 100, 65–79, 10.1016/j.jue.2017.05.001.

U.S. Department of Transportation (2009) *2009 National Household Travel Survey*, Washington, D.C., http://nhts.ornl.gov.

van den Berg, Vincent and Erik T. Verhoef (2011) "Winning or Losing from Dynamic Bottleneck Congestion Pricing?: The Distributional Effects of Road Pricing with Heterogeneity in Values of Time and Schedule Delay," *Journal of Public Economics*, 95 (7–8), 983–992, 10.1016/j.jpubeco.2010.12.003.

Verhoef, Erik T. (1999) "Time, Speeds, Flows and Densities in Static Models of Road Traffic Congestion and Congestion Pricing," *Regional Science and Urban Economics*, 29 (3), 341–369, 10.1016/S0166-0462(98)00032-5.

_____ (2001) "An Integrated Dynamic Model of Road Traffic Congestion Based on Simple Car-Following Theory: Exploring Hypercongestion," *Journal of Urban Economics*, 49 (3), 505–542, 10.1006/juec.2000.2203.

_____ (2005) "Speed-Flow Relations and Cost Functions for Congested Traffic: Theory and Empirical Analysis," *Transportation Research Part A: Policy and Practice*, 39 (7–9), 792–812, 10.1016/j.tra.2005.02.023.

Vickrey, William S. (1963) "Pricing in Urban and Suburban Transport," *American Economic Review*, 53 (2), 452–465, http://www.jstor.org/stable/1823886.

_____ (1969) "Congestion Theory and Transport Investment," *American Economic Review*, 59 (2), 251–260, http://www.jstor.org/stable/1823678.

——— (1987) "Marginal and Average Cost Pricing," in Durlauf, Steven N. and Lawrence E. Blume eds. *The New Palgrave Dictionary of Economics*, 1st edition, Basingstoke: Palgrave Macmillan, http://www.dictionaryofeconomics.com/article?id=pde1987_X001391.

Walters, Alan A. (1961) "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, 29 (4), 676–699, 10.2307/1911814.

Winston, C. and A. Langer (2006) "The Effect of Government Highway Spending on Road Users' Congestion Costs," *Journal of Urban Economics*, 60 (3), 463–483.

Zhang, Lei and David Levinson (2004) "Some Properties of Flows at Freeway Bottlenecks," *Transportation Research Record: Journal of the Transportation Research Board*, 1883, 122–131, 10.3141/1883-14.