

# Congestion Costs and Scheduling Preferences of Car Commuters in California: Estimates Using Big Data

Jinwon Kim\*      Jucheol Moon†

February 24, 2022

## Abstract

This paper aims to quantify congestion costs and estimate the scheduling utility function for commuters. To do so, we construct California commuters' travel-time profiles, namely, the menu of travel times that each individual will likely face according to alternate trip timing choices. On average, California commuters waste about 5 minutes per morning commute due to congestion. Commuters facing a higher congestion level at the peak hour tend to avoid congestion delays by arriving at an inconvenient edge time. We also discover that for the majority of the commuters in our data, travel-time profiles are much flatter than our estimated schedule utility. From this finding, we question the accuracy of the existing bottleneck models in quantifying the economic costs of congestion and the optimal toll to ameliorate congestion.

*Keywords:* congestion costs; scheduling preference; commuting; Google Maps; big data; machine learning

*JEL Classification Numbers:* R41, R48, C8, C25, H21

---

\*Department of Economics, Sogang University, Seoul, Korea. Email: [jinwonkim@sogang.ac.kr](mailto:jinwonkim@sogang.ac.kr).

†Department of Computer Engineering & Computer Science, California State University, Long Beach. Email: [jucheol.moon@csulb.edu](mailto:jucheol.moon@csulb.edu). SB1 grant from the California Department of Transportation, funded through the California State University Transportation Consortium, is greatly acknowledged. We thank David Brownstone, Kenneth Small, Filippo Tassinari, Erik Verhoef, and especially Jonathan Hall, as well as seminar participants at the 2021 ITEA, the 2021 KER International Conference, 2021 Urban Economics Association, and 2022 AREUEA-ASSA meetings, for helpful comments.

# 1 Introduction

Since [Pigou \(1920\)](#), scholars have sought to understand the phenomenon of traffic congestion. The growing number of recent empirical studies, which provide more reliable estimates on the congestion cost by using modern empirical tools and new data, confirm that the congestion cost is substantial (e.g., [Couture et al., 2018](#); [Yang et al., 2020](#); [Bento et al., 2021](#); [Russo et al., 2021](#)). However, these works usually focus on just one element of congestion models, particularly the slope of the road cost function, which is perhaps insufficient as a detailed policy guide toward congestion toll that would be imposed on real-world drivers who are heterogeneous in multiple dimensions.

The recent availability of real-time traffic information—namely “big data” like Google Maps and the development of prediction tools based on artificial intelligence, usually known as “machine learning”—is enabling researchers to explore the problem of congestion with a more structural framework. Among others, [Akbar and Duranton \(2017\)](#) and [Kreindler \(2020\)](#) have taken initial steps in this direction.<sup>1</sup> The current paper is in line with these recent papers, but ours is more closely tied to [Vickrey \(1969\)](#), which has become the workhorse model for analyzing scheduling behavior and congestion dynamics in economics (e.g., [Arnott et al., 1990, 1993](#)).

Specifically, this paper develops a simple model of trip scheduling under congestion to formally define the congestion delays the commuter faces and identify how commuters adapt to the congestion level they face (e.g., by adjusting route or timing choices). We then empirically quantify the model’s key elements by utilizing real-time travel time information from Google Maps and machine learning as a prediction tool. Our theoretical model features a salient fact that has mostly been ignored in the classical bottleneck models—individuals live and work at different locations and therefore each faces their own travel-time profile. Specifically, the commuter in our model chooses her optimal trip timing to minimize commuting costs subject to the travel-time profile faced on her commuter route as a constraint. In this situation, there may be a unique optimal trip timing on her commute route, which, in turn, implies that the shapes of congestion dynamics faced by individuals matter in their scheduling choices. Specifically, commuters facing a higher congestion level during the peak hour would choose an inconvenient early or late timing to avoid an otherwise longer congestion delay—a

---

<sup>1</sup>There are other studies of mobility and congestion using big data, especially in developing countries (e.g., [Akbar et al., 2019](#); [Kreindler and Miyauchi, 2020](#)). See [Selod \(2021\)](#) for a literature review on the uses of big data in transportation research more broadly. Also, for a survey of big data uses in urban economics research, see [Glaeser et al. \(2018\)](#).

finding for which we find statistically significant evidence.

The two key elements of the model are the travel-time profile faced by an individual and the scheduling utility, which jointly reveal the commuter's optimal arrival time choice. We empirically quantify both. A travel-time profile is the menu of travel times that an individual commuter faces on her commute route by alternate trip timing choices. We construct the systematic travel time predictions by alternate arrival time intervals, that is, travel-time profiles, for a large survey sample of commuters in California by querying travel times at Google Maps for each origin-destination pair with random timing in each day over one and half-year. Our travel survey dataset consists of 9,127 different zip code pairs of origin and destination, which are traveled by 14,544 commuters. This implies that a zip code pair and the corresponding travel-time profile is fairly unique to an individual commuter.

An especially useful quantity directly drawn from each constructed travel-time profile is the commuter's congestion delay (or "queuing time," borrowing from the bottleneck model's terminology), which is defined by the gap between the commuter's actual travel time and the counterfactual congestion-free travel time that would have been realized if she had traveled during times when there was no congestion. The classical bottleneck model (Vickrey, 1969; Arnott et al., 1990, 1993) suggests that this congestion delay is a pure social loss, and the aggregation of these losses equals the economic inefficiency costs from congestion (see also Kim, 2019). Our constructed travel-time profiles exhibit an average congestion delay of about 5 minutes per trip (about 20% of the sample mean of travel time), which implies inefficiency costs borne by morning commuters in California of about USD 6.6 billion.

The scheduling utility estimated in this paper comprises a travel time cost and schedule-delay costs that arise from arriving earlier or later than the commuter's ideal arrival time. The estimated parameters in the utility function reveal commuters' willingness to pay for a schedule delay reduction expressed in the unit of travel times. The first step toward estimating these parameters is to assign an ideal arrival time to each commuter, as it directly defines schedule delays as attributes in our random utility framework. For this, we apply the machine learning method to predict the counterfactual arrival time in the absence of queuing, learning from a group of commuters who are plausibly assumed to arrive at their own ideal arrival times. We update our machine-learning estimates on ideal arrival times using our theoretically grounded conjecture combined with the

results from the previous studies to finally estimate the structural parameters in the scheduling utility form.

Our scheduling utility estimates indicate that commuters are willing to accept about 0.5 additional minutes of schedule delay to reduce travel time by 1 minute. While the shape of the estimated scheduling utility is similar to those presented in previous studies (Small, 1982; Kreindler, 2020), we discover that it is much steeper than the travel-time profiles faced by commuters. Because travel time falls too slowly to compensate for the corresponding increase in schedule delay costs, this situation implies the tendency of commuters to choose an arrival time that is close to their ideal arrival times. Hall (2021b) discovered this key finding that travel time climbs and decreases too slowly relative to the indifference curve, pointing out the poor fit of the bottleneck model to travel time data. He suggested incorporating inframarginal travelers, defined as those strictly preferring to arrive at their ideal times, into the model to increase its fit to the data. Our solution is to incorporate the impacts of congestion on parts of commuters' routes other than a single point of congestion in analyzing commuters' scheduling choices. We also directly estimate individuals' ideal arrival times and find that they are quite dispersed, which explains why travel-time profiles are flat. Our approach can make a valuable start on a long-standing weakness of practical uses of the bottleneck model.

This paper adds to the literature estimating motorists' scheduling preferences and the distribution of preferences. In his seminal work, Small (1982) formulated the scheduling utility, the so-called  $\alpha - \beta - \gamma$  preference still used widely in the literature, and estimated these parameters. Peer et al. (2015) decomposed the morning scheduling decisions of car commuters into long-run choices of arrival routines and short-run choices of departure times subject to the routines chosen in the long run to estimate both short- and long-run preferences.<sup>2</sup> Kreindler (2020) estimated the scheduling utility for drivers in Bangalore, India, using, like us, Google Maps to construct the menu of travel times and adopting an experimental setting to explicitly estimate the parameters in terms of a monetary unit. Hjorth et al. (2015) estimated scheduling preferences using stated preference data incorporating non-linearity and travel time variability. Small et al. (2005) estimated the values of travel time and reliability as well as the distribution of those values. Hall (2021a) estimated the

---

<sup>2</sup>Verhoef (2020) investigates whether incorporating the empirically validated difference between long- and short-run preferences affects optimal pricing of congested roads.

joint distribution of scheduling preference parameters as well as ideal arrival times to evaluate the effects of an optimal time-varying tolling policy. There are also papers estimating the preferences of public transit users using revealed preference data (Peer et al., 2016; Hörcher et al., 2017). Our current paper uses a much larger sample of commuters to estimate the parameters more accurately and applies a machine-learning method to predict commuters' ideal arrival times in a scheduling utility form that is allowed to be non-linear.

Our paper is also closely related to the recent empirical papers estimating congestion costs. Most of these papers have focused on estimating Pigouvian deadweight loss by estimating the cost function of trips at the road or city levels (Walters, 1961; Couture et al., 2018; Yang et al., 2020; Russo et al., 2021). Meanwhile, Akbar and Duranton (2017) used big data to identify the demand curve as well as the cost curve to estimate the deadweight loss from congestion more accurately. Recent papers such as Tang (2021) and Tarduno (2021) adopted a quasi-experimental design to estimate congestion costs. Our approach to the estimation of congestion costs is most similar to that of Kim (2019), who also defined the congestion delay as the difference between the observed travel time and the counterfactual free-flow travel time to measure the economic cost of congestion defined in the standard bottleneck model (Vickrey, 1969; Arnott et al., 1990, 1993). However, unlike Kim (2019), who used only travel survey data and thus suffered from the identification problem, this paper observes counterfactual times at all hypothetical arrival time choices using Google Maps, which allows us to directly identify the congestion delay for each traveler.

The rest of the paper is organized as follows. Section 2 presents a model of trip scheduling under individual-specific travel-time profiles and develops an empirical framework for measuring the congestion delays of individual commuters. In Section 3, we explain the data. In Section 4, we examine the overall shapes of travel time and congestion delay profiles and quantify congestion costs. In Section 5, we estimate the causal effect of congestion on trip scheduling. In Section 6, we estimate the structural scheduling preference parameters. In Section 7, we discuss some implications of our findings. Finally, Section 8 concludes.

## 2 The Conceptual Framework

## 2.1 A model of commute scheduling

This section presents a simple model of trip scheduling for a representative commuter. Every commuter travels their own route to work simply because commuters all live and work in different places. Each commuter therefore faces her own menu of travel times according to alternate trip timing choices (i.e., the individual-specific travel-time profile).

The commuter's goal is to minimize commuting cost by choosing an arrival time, subject to the travel-time profile as a constraint. Commuting cost—with an arrival time  $t$ , denoted by  $C(t)$ —is assumed to have the following form:

$$C(t) = \alpha T(t) + \beta SDE(t)^{\omega_1} + \gamma SDL(t)^{\omega_2}, \quad (1)$$

where  $T(t)$  is travel time for the arrival time of  $t$ . With  $t^*$  indicating the commuter's ideal arrival time,  $SDE \equiv \max(t^* - t, 0)$  is schedule delay for arrivals earlier than  $t^*$ . The third term is the schedule-delay cost for arrivals later than  $t^*$ , with  $SDL \equiv \max(t - t^*, 0)$ . The parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  are the unit costs placed on each component. The commuter's departure time is simply  $d \equiv t - T(t)$ . Hence, under the assumption that the commuter is fully informed about his travel-time profile, the commuter effectively chooses an arrival time ( $t$ ) by choosing the corresponding departure time ( $d$ ).

Note that while the exponent of  $T(t)$  is normalized at unity, the exponents placed on the schedule delays satisfy  $\omega_1 \geq 1$  and  $\omega_2 \geq 1$ . It is useful to draw an iso-cost curve on the plane of arrival time  $t$  against travel time  $T(t)$  to appreciate the meaning of this assumption (see Figure 1). The iso-cost curve in the figure is drawn with the assumption that  $\omega_1$  and  $\omega_2$  are strictly greater than one. Because the second derivatives of the iso-cost curves are negative under  $\omega_1 > 0$  and  $\omega_2 > 0$ , the iso-cost curve is convex, i.e., gets steeper as  $t$  deviates more from  $t^*$ .<sup>3</sup> In other words, the required travel time reduction to maintain the cost from an increase in schedule delay increases as  $t$  deviates more from  $t^*$ .

When  $\omega_1 = \omega_2 = 1$ , our cost form is the same as the customary  $\alpha - \beta - \gamma$ , the preference of Small (1982) also used in Arnott et al. (1990, 1993), under which the marginal rate of substitution

---

<sup>3</sup>Differentiate (1) with respect to  $t$  and set it at zero to have  $T'(t)|_{\bar{C}} = (\beta/\alpha)\omega_1(t^* - t)^{\omega_1 - 1}$  for  $t < t^*$  ( $\bar{C}$  is a constant cost) and  $-(\gamma/\alpha)\omega_2(t - t^*)^{\omega_2 - 1}$  for  $t > t^*$ . The second derivative is  $T''(t)|_{\bar{C}} = -(\beta/\alpha)\omega_1(\omega_1 - 1)(t^* - t)^{\omega_1 - 2} < 0$  for  $t < t^*$  and  $-(\gamma/\alpha)\omega_2(\omega_2 - 1)(t - t^*)^{\omega_2 - 2} < 0$  for  $t > t^*$ .

between schedule delays and travel time is constant at  $\beta/\alpha$  before  $t^*$  and  $\gamma/\alpha$  after  $t^*$  (see Figure 2).<sup>4</sup>

Figure 1 about here.

Let  $T^P(t)$  be the travel-time profile *faced* by the commuter.  $T^P(t)$  is the expected travel time conditional on the commuter's arrival time  $t$ . With  $t_0$  being the time at which congestion sets in and  $t_1$  being the time at which congestion ends, it is assumed that  $T^P(t)$  is monotonically increasing from  $t_0$  until  $t^{peak}$  and decreasing from  $t^{peak}$  to  $t_1$ . The congestion-free travel of the commuter,  $T^P(t_0) = T^P(t_1) \equiv T^{free}$ , is the travel time if the commuter had traveled earliest or latest among her possible choices (an earlier arrival than  $t_0$  or a later arrival than  $t_1$  cannot be optimal if  $t_0 < t^* < t_1$ ).

We can also define a congestion-delay profile, which represents the congestion dynamics *faced* by the commuter. We denote it by  $Q(t) \equiv T^P(t) - T^{free}$ . The congestion delay is zero for an arrival of either  $t_0$  or  $t_1$  (i.e.,  $Q(t_0) = Q(t_1) = 0$ ). The maximum congestion delay on the commute route is  $Q^{peak} \equiv Q(t^{peak}) = T^P(t^{peak}) - T^{free}$ , where  $t^{peak}$  is the individual's arrival timing for which his travel time would be the longest.

To characterize the commuters' choice, consider, as an example, a bell-shaped travel-time profile, reflecting the real-world observations, which is illustrated in Figure 1. At any given arrival time choice  $t$ , any point below the travel-time profile is not feasible. So, at the cost-minimizing arrival time, the travel-time profile and an iso-cost curve meet while the iso-cost curve is placed at the lowest position possible. In the figure,  $t^c$  is the optimal arrival time chosen by the commuter. The following proposition characterizes optimal arrival time choice under two particular cases: for the proof, see Appendix A.

**Proposition 1** *Assume  $t_0 < t^* < t_1$  and  $T^P(t)$  is twice differentiable with  $T^{P''}(t) \leq 0$ .*

1. *Consider the case where the iso-cost curves have more negative second derivatives than  $T^P(t)$  globally over  $t \in [t_0, t_1]$ . In this case, there exists a unique optimal arrival time,  $t^c$ , between  $t_0$  and  $t_1$ . At  $t^c$ , the iso-cost curve and the travel-time profile are tangent.*

---

<sup>4</sup>Other studies use a more general formulation by imposing non-separability of arrival time and travel time. Under that condition, the optimal trip timing depends on the trip duration (see Fosgerau and de Palma, 2012; Fosgerau et al., 2018; Fosgerau and Kim, 2019).

2. Consider the case where the iso-cost curves have less negative second derivatives than  $T^P(t)$  globally over  $t \in [t_0, t_1]$ . In this case, the optimal arrival time  $t^c$  is  $t_0$  or  $t_1$  or  $t^*$ .

Proposition 1 suggests that each commuter has her own optimal trip timing most suitable to the travel-time profile she faces, which is a feature departing from the classical bottleneck model, where commuters are indifferent with alternate trip timing choices in the equilibrium. Note that Proposition 1 does not cover all cases, since the iso-cost curve may be only locally more or less convex than the travel-time profile. Our aim here is first to highlight the fact that the commuter adjusts her schedule to the travel-time profile of her own, rather than fully characterizing trip scheduling choices under all possible travel-time profiles.

Still we can discover adaptation patterns of commuters from the two cases examined in Proposition 1. We will specifically show that commuters facing a steeper travel-time profile is more likely to choose a non-peak time to avoid otherwise a long congestion delay. To see this, consider the first case where iso-cost curves are more convex (have more negative second derivatives) than  $T^P(t)$ . In this case, given that the condition for optimal choice is the equal slope of the travel-time profile to the slope of the iso-cost curve, if the former gets larger at the optimal point, then an extra schedule delay would decrease the travel time more than the required fall in travel time for the constant cost, inducing a choice that is farther from  $t^*$  and closer to  $t_0$  or  $t_1$ . This implies that commuters facing a steeper travel-time profile would tend to choose a time that is more distant from  $t^*$  and closer to  $t_0$  or  $t_1$ , while those facing a flatter travel-time profile would tend to choose a time closer to  $t^*$ .

Figures 2 and 3 about here.

In Figure 2, we illustrate the other case where the iso-cost curves have a smaller curvature overall than the travel-time profile, specifically by imposing the linear iso-curves with  $\omega_1 = \omega_2 = 1$ . In panel (a), we assume that the commuter faces a lower congestion level with a flat travel-time profile, according to which the commuter chooses to arrive at her ideal arrival time. In panel (b), the commuter is facing a higher congestion level. In this scenario, the commuter chooses an inconvenient edge time largely different from  $t^*$ .

Based on these observations, the effect of the travel-time profile on scheduling choices can be stated as follows:

**Corollary 1.1** *An increase in the slope of the travel-time profile induces an earlier arrival of the commuter if  $t^c < t^*$  and a later arrival if  $t^c > t^*$ .*

This corollary states that commuters adapt to the congestion dynamics they face, specifically by arriving either earlier or later than the peak time when they face a higher level of congestion. In Section 5, we test this hypothesis. Because the peak congestion delay on the commute route, i.e.,  $Q^{peak}$ , measures the overall slope of the travel-time profile faced by the commuter, we regress the commuter’s scheduling choice on  $Q^{peak}$ .

An interesting observation is that a commuter facing a higher congestion level may have a shorter “realized” congestion delay than when facing an otherwise lower congestion level because the commuter could avoid the long queue by arriving at a off-peak time. For example, in Figure 2, the realized congestion delay (i.e.,  $Q(t^c) = T^P(t^c) - T^{free}$ ) is shorter in panel (b) than in panel (a), even though panel (b) illustrates a higher level of congestion with a higher  $Q^{peak}$ .

Figure 3 exemplifies the average travel-time profile faced by individuals and scatter plots of arrival time and travel time. It is based on two hypothetical individuals with similar preferences but facing different congestion levels. Based on Corollary 1.1, this scenario works from the assumption that the commuter facing a higher congestion level chooses an edge time, denoted by  $t_h^c$ , while the one facing a lower congestion level chooses a peak time,  $t_l^c$ . In the figure, the long-dashed curve is the average travel-time profile, which connects the average travel time values of these two commuters by arrival time. Meanwhile, the short-dashed curve in the figure connects the chosen combination of arrival time and travel time, exemplifying scatter plots on arrival times and travel times drawn from the data. The large difference between the two (long- and short) dashed curves is due to the commuters’ adaptation behavior shown in Corollary 1.1.

## 2.2 Measuring commuters’ congestion delays

Here we provide the framework to measure individual commuters’ congestion delays (interchangeably used with “queuing times”) empirically. With  $i$  being the commuter subscript, the individual-specific congestion dynamics faced by the commuter can be written as  $Q_i(t) = T_i^P(t) - T_i^{free}$ . With  $t_i^c$  indicating the chosen (observed) arrival time, the individual’s “realized” queuing time is  $Q_i(t_i^c) \equiv T_i^P(t_i^c) - T_i^{free}$ . Travel diary data mostly report the travel time outcome corre-

sponding to  $T_i^P(t_i^c)$ , but not the congestion-free travel time  $T_i^{free}$ , which is the counterfactual travel time that would obtain if the traveler had alternatively chosen to travel earliest (or latest) in the morning and encountered no queues.<sup>5</sup> We overcome this identification challenge by constructing each commuter’s congestion-delay profile (or queuing-time profile) using the data on hypothetical trips queried via Google Maps.

We first construct the individual-specific travel-time profiles. Let  $\widehat{T}_i^P(m)$  be the systematic prediction of travel times for the origin and destination pair, namely the “route,” traveled by commuter  $i$ , where  $m$  is the set of alternate arrival time intervals, with  $m \equiv \{1, 2, 3, \dots, M\}$ . The predicted travel time must be “systematic” in the sense that the prediction rules out any day-specific unanticipated shocks, such as weather, traffic accidents, and roadwork, because commuters would make scheduling choices based on their systematic travel-time profiles.<sup>6</sup> To this end, we queried travel times for a long time horizon from Google Maps and then averaged out the travel time outcomes to construct the systematic travel-time predictions.

To construct the individual’s congestion-free travel time, i.e.,  $T_i^{free}$ , we use travel time predictions conducted during the lockdown policy under COVID-19, especially counterfactual trips with pre-6:00 AM arrivals queried between March 19 and June 30, 2020. We denote this by  $\widehat{T}_i^{free}$ . The estimate for the individual-specific congestion delay profile is then written as  $\widehat{Q}_i(m) \equiv \widehat{T}_i^P(m) - \widehat{T}_i^{free}$ . The “realized” delay of commuter  $i$  is obtained by evaluating it at the chosen arrival time, which is written as

$$\widehat{Q}_i(m_i^c) = \widehat{T}_i^P(m_i^c) - \widehat{T}_i^{free}, \quad (2)$$

where  $m_i^c$  is the arrival time interval containing  $t_i^c$ , i.e., the arrival time interval chosen by commuter  $i$ . A small number of commuters traveling uncongested routes have a negative value for  $\widehat{Q}_i(m_i^c)$ . We assume that their congestion delay is zero.

---

<sup>5</sup>It is useful to see how the counterfactual framework (see Angrist and Pischke, 2009) may be adopted to define the individual’s congestion delay alternatively. Assume that there are only two alternate trip timing choices: a time when there is congestion and another time when there is no congestion. Let  $T_{1i}$  be the travel time outcome if the traveler chose to arrive at a time with congestion and  $T_{0i}$  be the time if the traveler chose to travel at a time with no congestion (very early or very late in the morning). The congestion delay of the commuter is  $T_{1i} - T_{0i}$ , which is the *causal* effect of trip timing on travel time (see Kim, 2019). The traveler cannot choose both, so researchers can only observe either of them and thus cannot identify the congestion delay directly using travel diary data.

<sup>6</sup>Indeed, Peer et al. (2015) showed that commuters would form “routine” in response to their systematic travel-time profiles and then make daily choices on trip timing.

Another quantity estimated below is the peak queuing time of individual  $i$ , written as follows:

$$\widehat{Q}_i^{peak} = \widehat{T}_i^P(m_i^{peak}) - \widehat{T}_i^{free} \quad (3)$$

which evaluates the congestion delay profile at the interval  $m_i^{peak}$ , in which the travel-time profile of the individual  $i$  reaches its peak ( $m_i^{peak}$  contains  $t^{peak}$ ). This expression gives the expected queuing time conditional on arrival in the peak congestion condition of the commuter's own route. For commuters with a negative  $\widehat{Q}_i^{peak}$ , we again assume zero congestion.

In the empirical part, we first measure the congestion delays of a sample of commuters in California, i.e.,  $\widehat{Q}_i(m_i^c)$ , and use them to calculate congestion costs in California in Section 4. We then test the “adaptation” hypothesis of Corollary 1.1. Finally, we estimate the scheduling utility function to draw the iso-cost curves, which, together with the travel-time profiles, would reveal actual and ideal arrival times. As we will see, commuters in the real world have much steeper iso-cost curves than their travel-time profiles, as illustrated in panel (a) in Figure 1. Based on this observation, we draw a number of implications.

### 3 Data

The data on travelers' scheduling choices are obtained from the 2012 California Household Travel Survey (CHTS), which is a statewide survey conducted every 10 years. From the entire CHTS sample, we select morning commutes made by passenger vehicles to construct the estimation sample. We further limit our attention to morning commutes whose arrival times are between 5:00 AM and 11:15 AM. We exclude trips taken using public transit or ride sharing. We select this sample because the choice set in this sample is relatively clear, so it is easier to estimate the scheduling preferences. Also, as our model is closely tied to the Vickrey (1969) bottleneck model, we can test the power of this model by focusing on this sample.

An important information we obtain from the CHTS sample is the zip codes of the home and work locations of each commuter. The geographic centers in each zip code (for home and work) serve as the origin and the destination when we query from Google Maps. In our final sample of 14,544 commuters, there are 9,127 different zip code pairs (with 1,462 different zip codes for

home or work sites). So, although each zip code pair is not entirely unique to individuals (because commuters residing at different addresses may travel to or from the same zip codes), we can say that zip codes provide fairly detailed information on the origin and the destination of the commuters' trips.<sup>7</sup>

For each pair of zip codes ("route"), we construct the travel-time profile (i.e., the menu of travel times by commuter's alternate arrival time choices). For this aim, we queried over 18 million hypothetical trips from Google Maps in the period from January 6, 2020 to July 28, 2021, of which about 1.3 million are in our primary analysis period from January 6 to March 18, 2020 (the date just before the first lockdown policy enacted by the California government due to COVID-19 on March 19). The data collected after March 18 are excluded in travel-time profile construction due to the steep declines in traffic volume following the pandemic, but they are used to construct congestion-free travel times.

To construct each route's travel-time profile, we first classify the arrival times between 5:00 AM to 11:15 AM into 16 ( $= M$ ) intervals, so that each interval spans 15 minutes, to ensure that each interval contains modes of arrivals such as 7:00 AM, 7:30 AM in the center. The first and the last intervals (with  $m = 1$  or  $m = 16$ ) are set to be much wider, so these intervals cover 5:00 AM–6:22 AM and 9:52 AM–11:15 AM, respectively, to reflect low arrival rates and slowly changing traffic conditions during these times.

To measure the systematic prediction of travel time by interval, i.e.,  $\widehat{T}_i^P(m)$  for  $m = 1, 2, \dots, 16$ , we calculate the mean of predicted travel times from trips queried during our primary analysis period. In this way, we can rule out day-specific shocks such as accidents and weather conditions. Once we constructed travel-time profiles for each of 9,127 different pairs of zip codes, we then matched the profiles to individual commuters so that each has their own travel-time profile. Our final sample consists of 14,544 commute observations.

Finally, note that we must also assume that travel-time profiles constructed between January and March, 2020 are similar to those of 2012 when the commuters in the CHTS sample made their scheduling choices. Google Maps queries are always real-time, so we cannot go back and estimate the travel time that 2012 commuters would have experienced at that time. Because congestion

---

<sup>7</sup>For more information on how we identify routes for commuters traveling within the same zip code, see Appendix B.

is determined by long-term demand and supply factors, the cross-sectional variation in congestion across zip codes should remain stable. We provide further remarks on the data construction in Appendix B.

Table 1 about here.

Table 1 shows descriptive statistics for the key variables, such as arrival time, travel time, and trip distance. The arrival time variable ( $t$ ) is in minute and follows the decimal system with midnight being normalized at 0. We mainly use the arrival time variable, rather than the departure time, since the departure time suffers from a large heterogeneity in its values of commuters traveling difference distances even when they have a similar preference. The sample mean for arrival time (in minutes from midnight) is 478, i.e., 7:58 AM ( $478 = 7 \times 60 + 58$ ). The mean travel time is 25 minutes and the mean trip distance is 12 miles according to the CHTS sample. Google Maps predicted travel times for the commuters— $\widehat{T}_i^P(m_i^c)$  from (2)—have a sample mean of 26. The mean distance from Google Maps is a bit longer than the CHTS sample, implying that people tend to travel distances shorter than the distance between the central locations of the zip codes reported in the CHTS. The average speed (distance divided by travel time) from Google Maps is around 33 miles per hour.

Figure 4 provides scatter diagrams to visualize relationships between variables. From panel (a), we find that travel times are overall constant over the entire morning time interval. However, panel (b) shows that the speed (ratio of distance to travel time) tends to decline as the arrival time becomes later. Panel (c) shows a monotonic relationship between trip distance and travel time. Panel (d) shows that the speed is overall higher for longer-distance commutes in part due to the higher congestion level for commuters residing closer to their jobs (see Couture et al., 2018; Fosgerau and Kim, 2019).

Figures 4, 5, and 6 about here.

Figure 5 shows scatter diagram showing the fitness of the Google Maps predicted values to the CHTS reported values. In each figure, the horizontal axis represents the CHTS reported values, and the vertical axis plots the predicted value matched to each CHTS observation. As shown in panel (a), Google Maps data match quite well with the survey respondents' reported travel times,

but there are some errors—possibly because the survey respondents’ perceptions of their travel times are not always accurate. Panel (b) shows that the distance values in the two datasets match better than the travel times. The correlation coefficient for the travel time variables in the two datasets is 0.62, and for the distance variables it is 0.78.

Finally, Figure 6 illustrates the long-run evolution of travel times faced by California commuters during the pre-COVID-19 pandemic period and subsequent periods. Each point on the curve indicates the monthly cross-sectional average of Google Maps predicted travel times on the condition that commuters arrive in the peak hour (7:52 AM–8:52 AM).<sup>8</sup> As shown in Figure 6, in January and February 2020, the average of expected travel times conditional on arrival in the peak hour is around 28–29 minutes, but it dropped significantly when the statewide lockdown policy was enacted on March 19, 2020. While there was some recovery in travel time during the summer seasons of 2020 and 2021, travel times remain significantly low compared to the pre-COVID-19 values.

## 4 Congestion Costs

### 4.1 Travel time and congestion delay profiles

By taking the average of travel times of counterfactual trips whose arrival times fall in each interval  $m$ , which were queried between January 6 to March 18, 2020, we construct the individual-specific travel-time profile denoted by  $\widehat{T}_i^P(m)$ . This is the travel time that individual  $i$  would face if her arrival time belonged in the interval  $m$ . By arrival time interval  $m$ , we calculate the average of expected travel times from the commuters in our sample, which is written as

$$\overline{\widehat{T}^P}(m) \equiv \frac{\sum_{i=1}^N \widehat{T}_i^P(m)}{N}, \quad (4)$$

where  $N$  is the size of the sample ( $= 14,544$ ). This quantity is calculated for each  $m = \{1, 2, 3, \dots, M\}$  and plotted as the solid curve in Figure 7. Because we use the same sample to calculate the average expected travel time consistently for all intervals, we can interpret this curve as the average travel-time profile, or the travel-time profile of the commuter facing an average congestion level.

---

<sup>8</sup>Specifically, we first calculate each individual’s expected travel time if she arrived in the peak hour (7:52 AM–8:52 AM) by taking the mean of travel times from Google Maps queried trips whose arrivals belong in the peak time interval over each month. These expected peak travel times are then averaged out from the commuters in the sample with the size of 14,544, which yields each point on the curve.

According to the drawn curve, a typical commuter would spend around 23 minutes if she chose to arrive before 7:00 AM, and travel time is expected to be longer than 28 minutes if she can expect to arrive around the peak time, 8:15 AM.

Figure 7 about here.

The overall travel-time profiles that form *as a result of commuters' arrival time choices* are much different from the profiles *faced by* commuters. To show this point, we calculate the mean predicted travel time for each arrival time interval using only commuters who chose to arrive in that interval, which is expressed as

$$\tilde{T}(m) \equiv \frac{\sum_{i=1}^N \widehat{T}_i^P(m|m = m_i^c)}{N_m(m)}, \quad (5)$$

where  $N_m(m)$  is the number of commuters arriving in  $m$ . This is plotted by  $m$  and illustrated by the dashed curve in Figure 7. For example, 1,155 commuters arrived in the interval of 7:37 AM–7:52 AM ( $457 < t \leq 472$ ). The mean of their expected travel time is 26.7, etc. As can be seen in the figure, the travel-time profile based only on the chosen arrivals is flatter than the profile faced by commuters, which is attributed to the fact that commuters tend to adjust their arrival time choices to the travel-time profiles they face. Specifically, commuters traveling a longer distance and thus facing a long travel time at the peak would plan to arrive at a non-peak time. Because these travelers tend to arrive early in the morning, the mean travel time for early arrivals is quite high.

Figure 7 also shows how travel-time profiles changed from the travel demand shocks from COVID-19. In particular, we consider four sub-periods under the COVID-19 condition and draw mean travel-time profiles for each of them using Google Maps query outcomes drawn from each sub-period.<sup>9</sup> As anticipated, the travel-time profile is flattest for trips occurring just after the state's lockdown policy. We see some recoveries of the travel times for the next sub-periods, although they are still much flatter than under the pre-COVID-19 condition.

Now let us turn to congestion-delay profiles. We can construct each individual's congestion-delay profile using Google Maps by subtracting the individual-specific congestion-free travel time from the individual's travel-time profile, i.e.,  $\widehat{Q}_i(m) = \widehat{T}_i^P(m) - \widehat{T}_i^{free}$ . To measure each individual's

<sup>9</sup>The sub-periods include March 19 – July 31 (2020), August 1 – November 30 (2020), December 1 – March 31 (2021), and April 1 – July 28 (2021).

congestion-free travel time for each route, i.e.,  $\widehat{T}_i^{free}$ , we use trips with pre-6:22 AM arrivals using queries conducted during the COVID-19 lockdown policy (between March 19 and June 30, 2020). We anticipate that the lockdown policy increases remote working and thus reduces physical car commuting, eliminating congestion for morning commute trips. As confirmed in Figure 6, expected travel times are significantly lower during these days than before COVID-19. Furthermore, by focusing on pre-6:22 AM arrivals, like in Kim (2019), we can safely choose only trips that have not encountered any queues. We take the mean of travel time predictions for these trips to construct  $\widehat{T}_i^{free}$  for each  $i$ . Note that  $\widehat{Q}_i(m)$  is the congestion-delay profile that is *faced* by the individual, which is distinguished from the *realized* queuing time (2) (the difference comes from whether the predicted travel time  $T_i^P(m)$  is interacted with the commuter's choice to arrive in that interval).

By the arrival time interval  $m$ , the mean of  $\widehat{Q}_i(m)$ 's is calculated from the commuters in the sample, which is expressed as

$$\bar{Q}(m) \equiv \frac{\sum_{i=1}^N \widehat{Q}_i(m)}{N}. \quad (6)$$

This expression is plotted by  $m$  as the solid curve in Figure 8. This curve describes the average shape of the queue profiles faced by individuals. It has almost the same shape as the travel-time profile because the congestion-delay profile for an individual is obtained directly from her travel-time profile with only a level adjustment.

Figure 8 about here.

Now let us define the queue time profile that results from commuters' scheduling choices. Specifically, we define the mean queuing time conditional on commuters who arrived in the interval  $m$  as the following:

$$\tilde{Q}(m) \equiv \frac{\sum_{i=1}^N \left( \widehat{Q}_i(m) | m = m_i^c \right)}{N_m(m)}. \quad (7)$$

This is plotted by  $m$  as the dashed curve in Figure 8. Notably, the dashed curve during the peak hours between 7:00 AM and 8:30 AM is placed below the solid curve during the same interval, implying that the mean of queuing times for trips that chose this interval is smaller than the mean of all potential queuing times during this interval. This finding implies that commuters who would meet a relatively longer queue tend to avoid this interval, which possibly validates Corollary 1.1.

Figure 9 explores the heterogeneity of congestion profiles by different groups of commuters. In panel (a), we draw the queuing-time profiles faced by individuals by distance groups. They clearly show that commuters traveling a longer distance experience a long queue, which is seen by the large gap in travel times between non-peak and peak time for the longer-distance groups. Panel (b) draws congestion profiles for commuters by county of employers. As anticipated, counties known to be highly congested, such as San Francisco and Los Angeles, have much steeper congestion profiles than counties such as San Diego and Ventura. To further explore the difference in congestion by county, we report the congestion ranking in Table A1, where the ranking is based on the mean of per-mile expected queuing time during the peak hour.

Figures 9 and 10 about here.

Finally, Figure 10 again visualize the adaptation pattern of commuters. The histograms in the figure illustrate how commuters' scheduling choices differ by congestion level. Panel (a) illustrates the distribution of arrival times for the relatively highly congested Los Angeles County and the less congested Ventura County. Panel (b) illustrates the distribution of arrival times for commuters whose trip lengths are longer or shorter than the median distance (11.78 miles). Panel (c) illustrates the distribution of arrival times for commuters whose maximum queue on their commute routes exceeds 30 minutes and that of the rest of the commuters. All the panels imply that commuters subject to a higher congestion level tend to avoid the peak time.

## 4.2 Calculating congestion costs in California

Here, we aggregate congestion costs from our sample to estimate overall congestion costs in the California population. We have estimated the individual's queuing time that is "realized" as a result of commuters' scheduling choices (i.e.,  $\widehat{Q}_i(m_i^c)$  from (2)), which is different from the congestion-delay profile that is *faced* by the individual that was denoted by  $\widehat{Q}_i(m)$  above. The distribution of  $\widehat{Q}_i(m_i^c)$  is illustrated in the second column in Table 2. The third column in the same table illustrates distribution of peak queuing times that they would face in the peak point (i.e.,  $\widehat{Q}_i^{peak}$ ).

Table 2 about here.

The total of realized queuing times from our sample and the average queuing time are expressed as follows:

$$TQ \equiv \sum_{i=1}^N \widehat{Q}_i(m_i^c), \quad AQ \equiv \frac{TQ}{N}, \quad (8)$$

where  $N$  is the number of observations in our sample. We find that  $TQ$  for all 14,544 commuters in our sample is 67,669 minutes, with  $AQ$  being 4.65 minutes, which is about 20% of the sample mean of travel times. The median queuing time is 2.08, which is much shorter than the mean queuing time because the queuing times are skewed toward longer queuing times.

We also calculate the weighted average of queuing times, using weights based on the proportion of the in-sample county population to the county population in California, to deal with the large heterogeneity in the congestion levels across counties and the different distribution of county observations from the population distribution. Specifically, we calculate

$$AQW \equiv \sum_{i=1}^N w_i \widehat{Q}_i(m_i^c), \quad \sum_{i=1}^N w_i = 1. \quad (9)$$

The weight placed for  $i$  is proportionate to the inverse of the proportion of the number of county observations in our sample to the county population. It is to compensate commuters from counties with fewer observations in our sample.<sup>10</sup> We find that the weighted mean of queuing times is about 5.14 minutes, which is a bit larger than the unweighted mean, implying an oversampling of commuters located in relatively less congested areas in the CHTS sample.

We use the weighted mean of queuing times to calculate the total congestion costs in California. We note that the labor force in California in July 2021 is about 18.9 million and that the percentage of workers who commuted by private vehicle (either driving alone or carpooling) is about 85.3%, and we conclude that about 16.1 million workers used cars for their regular commutes in that year.<sup>11</sup> We also assume that the number of workdays for typical workers in the US in a year is 220 days, although this number should vary by worker. We therefore multiply our estimate for queuing time (5.14 minutes) by 16.1 million workers and 220 (days) to conclude that the total travel time loss from congestion for the California car commuters in a year is about 303 million hours.

<sup>10</sup>Specifically, we first calculate the ratio of county sample size in our CHTS sample to the county population size and assign a weight to each county by the inverse of this ratio. We then aggregate them all over individuals, which we based on to assign an individual's weight to ensure that the sum of weights is one.

<sup>11</sup>The sources of the labor force and the percentage of driving workers are the US Bureau of Labor Statistics and the American Community Survey by the US Census Bureau.

To finally calculate the dollar value of the time loss, we need an estimate for the value of time (VOT). From an extensive survey, [Small and Verhoef \(2007\)](#) conclude that the value of time for personal journeys varies widely by circumstances usually between 20% and 90% of the gross wage rate and averaging around 50%. Since car drivers value time more highly under congested conditions than under free-flow conditions, by perhaps 25% to 55% (see [Small, 2012](#); [Wardman, 2001](#)), we conclude that 70% of the gross wage rate is most appropriate for the average VOT. With the average hourly wage in California in 2020 of USD 31 (sourced from the Bureau of Labor Statistics), the VOT we choose is therefore USD 21.7 per hour. It then yields a total congestion cost of about USD 6.6 billion from morning commutes in California. An important issue of whether our calculated congestion cost is equivalent to the economic cost of congestion remains, which we discuss in detail below.

## 5 The Effect of Congestion on Scheduling Choices

The key theme of this paper is that people all have different routes (paths) to work. A natural question would then be, how would an individual commuter respond to the condition of her specific route? Our theoretical results in [Section 2](#) suggests that a commuter facing a steeper travel time profile is more likely to choose an inconvenient edge time, which has been validated in a descriptive manner in the preceding section. In this section, we use a regression model to test this hypothesis of commuter adaptation to routes.

We particularly aim at estimating the causal effect of congestion on trip timing choice. [Corollary 1.1](#) suggests that commuters on a steeper travel-time profile tend to avoid the peak time. Since the overall slope of the profile is measured by  $\widehat{Q}_i^{peak}$ , which is the peak height of the congestion-delay profile, we regress the scheduling choice on  $\widehat{Q}_i^{peak}$  to test this hypothesis.

We estimate a simple model of binary choices between an arrival during the peak hour (7:52 AM–8:52 AM), an hour interval during which the level of congestion on average is highest) and non-peak hours. In our sample, about 29% of commuters arrive during this peak hour. Our estimating equation is written as

$$arrpeak_i = \delta_0 + \delta_1 \widehat{Q}_i^{peak} + X_i \Gamma + \varepsilon_i, \quad (10)$$

where  $arrpeak_i$  is an indicator whose value is 1 if arrival is in the peak hour (7:52 AM–8:52 AM),

$X_i$  is the vector of observable worker and trip characteristics, and  $\varepsilon_i$  is the error term. The adaption hypothesis implies a lower probability of choosing the peak hour if the commuter were to travel a more congested route, under which  $\delta_1 < 0$ . Importantly—unlike the realized queuing time (2)—the key explanatory variable in (10) is not a function of the arrival time itself (see (3)). So, there is no reverse causality issue, at least not by the construction of the variables themselves.

Table 3 reports the estimation results. First, from the simple linear regression reported in column (1), we find that the estimated coefficient,  $\delta_1$ , is negative, confirming that commuters meeting a longer queue at the peak tend to arrive at a non-peak time. In column (2), we control for the trip distance and still find a negative coefficient, implying that adaptation behavior is not merely sorting arrivals by trip distance. In column (3), we include a set of personal characteristics as controls. In column (4), we add the county dummies, the workers' occupation dummies, and the industry dummies. We find that the explanatory variable becomes stronger (yielding a more negative  $\delta_1$ ) when more controls are added.

Tables 3 and 4 about here.

We would like to be careful about the possibility of sorting, under which scenario our estimate may be biased. Specifically, a higher congestion level on a route at the peak hour may mean that the users of the route tend to have a stronger preference toward the peak hour, so a commuter's arrival at the peak hour may reversely explain the higher congestion level on her route. While we try to control for factors determining commuters' scheduling preferences, the concern of omitted variables remains.

We use the instrumental-variable (IV) regression to address this concern. We use the mean congestion of workers who have the same work zip code as the commuter as the IVs, specifically the mean of  $\widehat{Q^{peak}}$  of other workers who have the same zip code in column (5) and the mean of per-mile  $\widehat{Q^{peak}}$  of other workers who have the same zip code in column (6), respectively. These variables are expected to reflect the congestion level around the workplace of the commuter, not the scheduling preference of the commuter himself. We find that the IV estimates on  $\delta_1$ , reported in columns (5)–(6), are similar to the OLS estimates. Finally, column (7) reports the Probit estimation result, which exhibits a rather large effect of congestion on the peak-time arrival choice.

The adaptation tendency implied in our estimates is nevertheless small. For example, for an

8-minute increase (i.e., the sample mean for  $\widehat{Q^{peak}}$ ) in the peak queuing time, there is only about a 2% reduction in the probability of choosing the peak hour. As we shall argue later, the small adaptation tendency does not imply that commuters are very inflexible. Rather, our scheduling preference parameter estimates imply that commuters are moderately flexible (they are willing to delay about 1 hour either earlier or later than the ideal time to reduce travel time by 30 minutes). We will provide our explanation reconciling these findings below.

In Table 4, we estimate several alternative specifications to further test the adaptation hypothesis. First in column (1), we use only the sample of trips with arrivals earlier than 8:52 AM to observe the choice between an arrival at the peak (7:52 AM–8:52 AM) and an earlier arrival (than 7:52 AM), and we find a slightly stronger adaptation tendency. In column (2), we use only the sample commutes arriving at the peak hour and later times (7:52 AM–8:52 AM vs. later than 8:52 AM) and find a similar adaptation tendency as those in Table 3.

In column (3), we use the queuing time per mile as the explanatory variable, and we again find a statistically significant effect of congestion on the tendency to choose the peak hour. In column (4), we modify the queuing time to the log scale and find that the elasticity of choice probability with respect to log of travel is small. For a 100% increase in the queuing time at the peak, we find a 1.4% lesser chance of choosing the peak time as the arrival time. Finally, in column (5), we used the log of queuing time per mile, which exhibits a similar estimate for the elasticity.

Overall, we find a statistically significant evidence of adaptation behavior, consistent with the hypothesis from our conceptual framework. Specifically, commuters traveling a highly congested route during the peak hour tend to adapt to the condition by arriving at a non-peak time (i.e., congestion on a route causes commuters to adjust when they depart). However, the adaptation tendency is quite weak. For example, for a 100% increase in queuing time at the peak hour, there is only a 1–2% lower chance of choosing the peak time. For the small adaptation tendency, we need an explanation, which we provide below after estimating the scheduling preferences in the next section.

## 6 Estimation of the Scheduling Utility Function

In this section, we estimate the schedule preferences of California commuters. The knowledge on scheduling preferences is used to develop useful implications for tolling policy.

### 6.1 The utility formulation and the logit model

An individual is assumed to choose an arrival time interval among 14 intervals including  $m = 2, \dots, 15$ , with each interval spanning 15 minutes. The entire choice set of arrival times covers 6:22 to 9:52 AM. We exclude the trips with intervals of  $m = 1$  and 16 (i.e., times before 6:22 AM or after 9:52 AM) because these travelers would have quite different preferences. We first consider a linear form of scheduling utility. The systematic utility that the commuter  $i$  obtains by choosing arrival time interval  $m$  in this form is written as

$$V_{mi} = -\alpha T_{mi} - \beta SDE_{mi} - \gamma SDL_{mi} - \eta DUM15_m, \quad (11)$$

where  $T_{mi}$  is the travel time that  $i$  would face if choosing the interval  $m$ ;  $SDE_{mi} = \max(0, t_i^* - t_{mi})$ , where  $t_i^*$  is  $i$ 's ideal arrival time and  $t_{mi}$  is the mid-point time of the interval in which  $i$ 's arrival time belongs; and  $SDL_{mi} = \max(0, t_{mi} - t_i^*)$ . Importantly, individuals have all their own ideal arrival times  $t_i^*$ , so they face their own menu of travel times and schedule delays (these utility components involve person ID). In the specification, we also include  $DUM15_m$ , which indicates an arrival interval that ends on the 15th or 45th minute in each hour. It controls for the traveler's tendency of rounding off their arrival time reports around 30 minutes, so the frequency of arrivals ending on the 15th or 45th minute in each hour is systematically smaller.

We also estimate the quadratic utility of the following form:

$$V_{mi} = -\alpha T_{mi} - \beta_1 SDE_{mi} - \beta_2 SDE_{mi}^2 - \gamma_1 SDL_{mi} - \gamma_2 SDL_{mi}^2 - \eta DUM15_m. \quad (12)$$

Below, we also estimate specifications including interaction terms of schedule delays and individual characteristics to uncover the dependency of scheduling preferences.

With  $\epsilon_{im}$  being the individual-specific factors determining the utility, the utility is  $V_{mi} + \epsilon_{im}$ . In the multinomial logit framework, in which the extreme value distribution of  $\epsilon_{im}$  is assumed, the

probability of choosing alternative  $m$  from choice set  $M$  by decision-maker  $i$  is given by:

$$P_{mi} = \frac{\exp V_{mi}}{\sum_{l=1}^M \exp(V_{li})}, \quad (13)$$

and the log-likelihood is written

$$LL \equiv \sum_{i=1}^N \log(\Pi_{mi}^{y_{mi}}), \quad (14)$$

where  $y_{mi} = 1$  indicates that  $m$  is  $i$ 's chosen alternative (and  $y_{mi}$  is otherwise 0). The maximum likelihood estimation is used to estimate the values of the preference parameters, including  $\alpha$ ,  $\beta$  (or  $\beta_0$  and  $\beta_1$  in the quadratic specification), and  $\gamma$  (or  $\gamma_0$  and  $\gamma_1$ ).

## 6.2 Estimating ideal arrival times

### 6.2.1 Machine-learning estimation

It is essential to first estimate individual-specific ideal arrival times because they define the schedule delays as a key attribute considered in decision making. We use the machine-learning method to estimate individuals' ideal arrival times.

The first step is to define the “example group” as the group of commuters who would reveal how agents would behave in the absence of congestion. Meeting no congestion on their route, these commuters would arrive at their ideal arrival times. In the variants of the bottleneck model in which residents are distributed over space and travel different distances, including Fosgerau and de Palma (2012), Fosgerau et al. (2018), and Fosgerau and Kim (2019), commuters traveling the shortest distance do not meet congestion and thus tend to arrive at their most preferred arrival times. Motivated by this observation, we define the example group of commuters as the commuters who have the same zip code for home and work, with a distance under 3 miles and a self-reported trip duration of less than 10 minutes. The rest of the commuters are in the “learning group,” for whom the ideal times should be estimated. Of the 10,487 commuters in the sample for our discrete choice estimation, 1,134 are in the example group, and 9,353 are in the learning group.<sup>12</sup>

---

<sup>12</sup>Note that the sample size for discrete choice estimation is smaller than that used in the congestion costs calculation. It is because we do not include commuters who arrive at  $m = 1$  or 16, and the predicting variables in machine-learning estimation (e.g., incomes and job categories) are missing some commuters, so their ideal time estimates are missing.

Figure 11 about here.

In Figure 11, the solid curve is the distribution of the arrival times of the commuters in the example group, and the dotted curve is the distribution of arrival times of the “learning group”. That the arrival times of the commuters in the example group are more concentrated around the peak than those of the learning group is consistent with the theory such as Fosgerau and Kim (2019).

The “machine” first learns how observed characteristics, such as incomes, occupation, and family size, affect the arrival times of commuters in the example group. We find that the information on workers’ jobs (composed of 21 industry dummies and 24 occupation dummies) has an especially large contribution to the prediction of arrival times. Using the information learned from choices made by commuters in the example group, the machine then predicts counterfactual arrival times that would be chosen in the absence of queues (i.e., ideal arrival times) for the learning group. Because we assume that commuters traveling a short enough distance would arrive at their ideal times, we impose the average distance of the example group (1.4 miles) as the common trip distance for the learning group in predicting their counterfactual arrival times. Then, the machine learning effectively matches commuters between the example and non-example group based on observable characteristics to predict ideal arrival times conditional on the short trip distance.

Figure 12 about here.

It is important to check the prediction power of the machine-learning estimates. We carry an out-of-sample test, by which we randomly select 30% of observations in the example group and predict their arrivals using the rest in the example group (70%). Because the selected group of 30% of commuters have both predicted and actual values of arrival times, we can then check the prediction power. Figure 12 plots actual and predicted arrival times of the selected commuters. We find that machine learning tends to over-estimate the arrival times for earlier arrivals and under-estimate for later arrivals, so that the distribution of predicted ideal arrival times is systematically too concentrated around the center (see the dashed curve in Figure 11, which illustrates the distribution of the machine-learning estimates for the ideal times). This outcome is probably because early or late arrivals are not quite predicted by any systematic factors, so they are determined mainly by individual-specific factors.

## 6.2.2 Updating ideal arrival time estimates using conjecture

We seek to correct the systematic prediction errors of the machine-learning estimates found in Figure 12. We use a theory-based conjecture for this purpose. We note that the travel-time profiles constructed in this paper are quite flat, even for highly congested routes during rush hour. From the travel-time profiles drawn from Google Maps, the travel time on average increases only about 7 minutes during the first half of the morning period (from an early time, say 6:00 AM, to the peak time, 8:00 AM). Meanwhile, earlier studies estimating scheduling preferences (Small, 1982; Kreindler, 2020) report an estimated  $\beta$  to  $\alpha$  ratio of 0.3–0.5, implying a one-hour increase in travel time corresponding to two-hour changes in arrival times. So, the travel-time profiles should be much flatter than most commuters' iso-cost curves.

Figure 13 about here.

This situation is described in Figure 13. In the graph, we consider five different individuals facing a common travel-time profile. We hypothesize non-linear iso-cost curves, whose slopes are higher than the travel-time profile. We can draw important implications from this situation. In particular, given that the condition for an arrival at a non-preferred time (i.e., time that is largely different from  $t_i^*$ ) is a steeper travel-time profile than the iso-cost curve around  $t_i^*$  (see Section 2), the situation in Figure 13 implies this condition is hard to meet. So, commuters choosing a non-peak time would do so mainly because these non-peak times are actually close to their preferred times and not because they adapt to potential long delays at the peak hour. Moreover, given the large difference in slopes between travel-time profiles and iso-cost curves, a small increase in the slope of the travel-time profile would not be enough to induce a rescheduling of arrivals to a non-peak time, thus explaining the weak adaptation tendency found in Section 5.

Simply put, the commuters tend to arrive at a time quite close to their ideal arrival times. However, as machine learning does not utilize the commuters' actual arrival times in the prediction, the estimates on ideal arrival times would be overly different from actual arrival times. Furthermore, the systematic over- (under) estimation of ideal times for early (late) arrivals observed in Figure 12 implies that the naive machine-learning estimates are too concentrated around the peak hour relative to the distribution of actual arrival times.

We, therefore, weight the machine-learning estimates and actual arrival times in the following form to update ideal arrival time estimates and assign them to commuters:

$$\widehat{t}_i^* \equiv \pi \widetilde{t}_i^* + (1 - \pi)t_i, \quad (15)$$

where  $\widehat{t}_i^*$  is the updated estimate for  $t_i^*$ ,  $\widetilde{t}_i^*$  is the naive machine learning estimate,  $t_i$  is commuter  $i$ 's actual arrival time, and  $\pi$  is a parameter with  $0 < \pi < 1$ . Under the assumption that the ideal arrival time is close to the actual arrival time, the update tends to reduce the above-mentioned systematic prediction errors.<sup>13</sup> Note that our weighting approach may be interpreted in the other direction: first randomizing ideal times around the actual arrival times, and the generated errors then being corrected by incorporating individuals' observed characteristics using machine learning.

We are left with the choice of  $\pi$ . By choosing  $\pi$ , we effectively choose our assumption on the distributions of finally assigned ideal times (i.e.,  $\widehat{t}_i^*$ 's). If  $\pi = 1$ , then naive machine-learning estimates are fully used, so the distribution of ideal times would be concentrated around the peak hour (see the dashed curve in Figure 11). As we decrease  $\pi$ , since actual arrival times are more spread out, the distribution of  $\widehat{t}_i^*$  becomes more flattened. In choosing  $\pi$ , we first ensure that the distribution of ideal arrival times resembles the distribution of arrival times for the example group, while utilizing the machine-learning estimates with the use of a high  $\pi$  as much as we can. We also ensure that the estimated utilities (linear and quadratic forms) fit the data well. In Figure 11, the dash-dotted curve indicates the distribution of  $\widehat{t}_i^*$ 's with our chosen value  $\pi = 0.4$ . Because it remains to be questioned how our estimates depend on  $\pi$ , we test with varying  $\pi$  values.

### 6.3 Scheduling utility parameters estimates

Table 5 reports the preference parameters estimated based on our chosen parameter value  $\pi = 0.4$ . In the linear specification reported in column (1), we find that the estimated coefficient of  $T$  is  $-0.121$ , and that of  $SDE$  is  $-0.0622$ , so the ratio of  $\beta$  to  $\alpha$  is about 0.514. Since the coefficient

---

<sup>13</sup>To see this point, let  $e_i = t_i^* - \widetilde{t}_i^*$  be the error of the machine-learning estimate. Plugging this expression into (15) to eliminate  $\widetilde{t}_i^*$  and rearranging terms, we can write down the prediction error from using  $\widehat{t}_i^*$  as follows:

$$t_i^* - \widehat{t}_i^* = (1 - \pi)(t_i^* - t_i) + \pi e_i.$$

From Figure 12, we can see that for earlier arrivals than ideal times with  $t_i^* - t_i > 0$ ,  $e_i$  tends to be negative, and for late arrivals with  $t_i^* - t_i < 0$ ,  $e_i$  tends to be positive. Because  $0 < \pi < 1$  is assumed, the first and the second term therefore would tend to cancel out, so  $t_i^* - \widehat{t}_i^*$  would tend to be smaller than  $e_i$ .

of  $SDL$  is  $-0.066$ , the estimated ratio  $\gamma/\alpha$  is  $0.545$ . These numbers mean that commuters are willing to arrive early or late by an hour on average to reduce the travel time by 30 minutes. Or equivalently, the cost of staying at work early before one's shift starts (and the cost of being late) is about half of the time spent in the car.

The coefficient on  $DUM15_m$  is significantly negative in all specifications, implying that survey respondents round off their arrival times by around 30-minute intervals. In specification (2), we include the square terms of the schedule delays. The large and significant coefficients on the quadratic terms support the non-linearity of the utility, approximated by the quadratic form.

Table 5 about here.

In the specifications reported in columns (3)–(5), we include interaction terms to see how commuters in one group value the utility components differently from those in the other group. First, in column (3), we interact the utility components with the dummy for “rich” commuter, which indicates an annual income exceeding USD 100,000. We find that the higher-income group is a bit less sensitive to schedule-delays for an early arrivals while much more sensitive to late arrivals. The travel time cost  $\alpha$  is not significantly different by income group. The ratio of  $\beta$  to  $\alpha$  is higher for the low-income group while the ratio of  $\gamma$  to  $\alpha$  is higher for the high-income group. Because it feels more costly to arrive later than  $t^*$  for the higher-income earners, they would tend to arrive relatively earlier. In column (4), we include the interaction terms with the female dummy, showing that women are relatively more sensitive or inflexible to schedule delays than men. In column (5), we define a “flexible” worker dummy which indicates workers in the CHTS sample who said that they are fairly flexible in adjusting their work schedules or that their days at the primary workplace are fewer than five. We can anticipate that these workers are also flexible in adjusting their scheduling within a work day. As expected, flexible workers have relatively smaller schedule delay costs.

Table 6 about here.

Of course, our utility parameter estimates depend on  $\pi$  as it determines the distribution of ideal arrival times and thus schedule delays faced by commuters. To see this dependency, we estimate the linear utility specification with varying  $\pi$  and report the estimated ratios  $\beta/\alpha$  and  $\gamma/\alpha$  at varying

$\pi$ 's in Table 6. From the table, we can set a reasonable range for the marginal rate of substitution between schedule delay costs and travel time costs. We first confirm that the estimated ratios  $\beta/\alpha$  and  $\gamma/\alpha$  are less than 1 except for the cases where we used  $\pi = 0.1$  or  $\pi = 1$ . For a fairly wide range of  $\pi$  that is 0.3–0.8,  $\beta/\alpha$  is between 0.47 and 0.62, and  $\gamma/\alpha$  is between 0.49 to 0.66. The mean of  $\beta/\alpha$  from this  $\pi$  range is 0.51 and the mean of  $\gamma/\alpha$  is 0.55, which are similar to the estimates based on  $\pi = 0.4$ . It would be fair to say that the unit cost of schedule delays ( $\beta$  and  $\gamma$ ) is about half of the unit cost of travel times.

Here is a short retrospective discussion on our assumption to use equation (15), which was made from our theory-based conjecture. A presumption behind this conjecture was that the iso-cost curve is steeper than the travel-time profile, which may be unduly forcing the revelation of such a situation. However, even with a higher value for  $\pi$  (using less of the conjecture), or even with  $\pi = 1$  (which does not use the conjecture at all), our estimated parameters still imply steeper iso-cost curves. Thus, the conjecture itself does not drive the conclusion that the iso-cost curves are steeper, but rather, we use it to estimate the scheduling utility parameters more accurately.

## 7 The Implications of the Findings and Discussion

In this section, we discuss the implications of our findings and the bottleneck modeling tradition based on our empirical evidence. For this aim, it is necessary to first interpret the magnitudes of our estimates. In panel (a) of Figure 14, we use our parameter estimates on the linear utility to draw an indifference curve with the assumption that this commuter's ideal arrival time is 8:15, which is the peak point according to the average travel-time profile. The average travel-time profile of Figure 7 is also placed in this figure to examine the commuter's arrival time choice. Given  $\beta/\alpha$  (and  $\gamma/\alpha$ ) of about 0.5, travel time must fall by about half an hour to compensate for the increase in schedule-delay costs arising from arrivals one hour earlier or later than the ideal time. However, the much flatter travel-time profile means that travel time on average falls by at most 2–3 minutes during the one-hour interval. As a result, the commuter chooses to arrive at her ideal time, which is 8:15 AM in this example.

Figure 14 about here.

In panel (b), we use the quadratic utility estimates to draw an indifference curve. Again, travel time on average still falls too slowly to compensate for the increasing schedule delay costs from arrivals later or earlier than the ideal time, under which the optimal choice is to arrive at the ideal time. If we assume she has a different ideal arrival time, then under the convex utility, the commuter's chosen time may be a bit different from her ideal arrival time. However, as long as the indifference curve is much steeper than the travel-time profile, the gap will only be small. As a result, commuters would have a strong tendency to arrive near their ideal arrival times, although they have moderate schedule inflexibility costs.

With the large difference between the travel-time profiles and the indifference curves observed in Figure 14 (as well as in Hall (2021b)), the two main results from the bottleneck model, in which the travel-time profile and the indifference curve are the same in equilibrium, should remain unresolved. The first is on the quantification of the economic cost of congestion. We can tentatively narrow the scope of our discussion by assuming that, like in the standard bottleneck models, the social optimum involves no queues. In this case, if the travel-time profiles were the same as the iso-cost curves, then the total queuing times quantifying the difference in total travel times between the social optimum and the laissez-faire, equivalent to the inefficiency cost arising from traffic congestion, could be measured simply by evaluating the travel-time profiles at the chosen arrival times as we have done in this paper (see also Arnott et al., 1990, 1993; Kim, 2019). However, given the large difference between the travel-time profiles and the indifference curves, measuring the congestion cost becomes a complicated issue. The trade-offs implied in the two curves are different. In particular, the travel-time profile gives the schedule delays the commuter *must accept* to have a zero queuing time under the laissez-faire condition. The indifference curve gives the schedule delays she would be *willing to accept* for the removal of queuing she would otherwise experience under the laissez-faire condition. A difficult issue is which one to use when evaluating the economic benefit available from the removal of queues (or equivalently, when evaluating the cost from the presence of queues). This issue becomes even more complicated if social optimum involves some queues.<sup>14</sup>

The second issue is how one could design a tolling policy to achieve the social optimum in a

---

<sup>14</sup>Indeed, in the more traditional congestion models based on Pigou (1920), where the congestion externality is the source of market failure, there is some congestion in the social optimum. Based on this model, Duranton and Venables (2018) argue that the congestion-free time may not be the reference point for identifying the social optimum. The Pigou and the bottleneck model also differ in the congestion technology, particularly the assumption on the relationship between speed and traffic volume.

decentralized setting under the large difference between the travel-time profiles and the indifference curves. In the classical bottleneck model, because the travel-time profile and the indifference curve are the same in equilibrium, the optimal time-varying toll schedule could be either. However, our result suggests that neither may correspond to the socially optimal toll. Specifically, a toll resembling the travel-time profile (suggested in [Kim, 2019](#)) would not induce commuters to reschedule their trips because the maximum benefit from rescheduling is the full removal of queuing times (7–8 minutes at most for typical commuters). This benefit would not be enough to compensate the schedule-delay costs of several hours for those having average preferences. A utility-compensating toll (having the shape of the indifference curve) may impose an inefficiently large toll for an arrival around the peak time, given that arrival times are already quite dispersed.

In his recent study, [Hall \(2021b\)](#) confronts the same fact of travel times climbing and falling too slowly relative to the indifference curves. For the problem of a poor fit of the bottleneck model to his travel time data, he suggests the incorporation of “inframarginal” travelers, defined as those having exceptionally strict preferences toward their ideal times, so they arrive at their ideal times even with some change in the shape of the travel-time profiles. However, our finding suggests that the majority of commuters may be “inframarginal”, so adding a smaller number of these commuters in the model would not be enough to solve the problem.

We suggest fundamental modifications of bottleneck modeling. In particular, the bottleneck models could incorporate the salient but ignored fact that commuters each face their own travel-time profiles and adjust their schedules to them. We would extend this model by incorporating the large dispersal of ideal arrival times. The dispersion of ideal times can resolve the seemingly contradictory result, that commuters tend to arrive around their ideal times while the resulting travel-time profiles are quite flat.<sup>15</sup> However, it would still be challenging to fully endogenize the heterogeneous travel-time profiles because each commuter is just one person contributing to the formulation of an individual travel-time profile that is shaped by others who, in turn, are shaping their own travel-time profiles. A welfare analysis to solve the above-mentioned problems of the

---

<sup>15</sup>Note that incorporation the heterogeneity in the ideal arrival times is not new in the literature. Even the original bottleneck model of [Vickrey \(1969\)](#) assumed a uniform distribution of heterogeneous ideal arrival times (see also [Small and Verhoef, 2007](#), who explained the original bottleneck model with the uniformly distributed ideal times.). However, this feature of his model has mainly been eliminated from most of the literature. [Arnott et al. \(1994\)](#) also considered the preference heterogeneity of commuters in terms of ideal arrival times, but there is no heterogeneity in commuters’ travel-time profiles in that paper. Several empirical studies support large dispersion of ideal times (see [Small et al., 2005](#); [Anas, 2015](#); [Kreindler, 2020](#); [Hall, 2021a](#)).

bottleneck model would therefore still be challenging. The current paper perhaps makes a start on a long-standing weakness of practical uses of the bottleneck model to solve these ultimate problems.

## 8 Conclusion

In this paper, we developed a simple model of trip timing choices faced by commuters facing their own travel-time profiles. We empirically quantified the model's key elements, such as the travel-time profiles and the scheduling utility function of car commuters in California. Using these parameters, we quantified the total congestion costs from morning commutes in California. We also empirically tested a hypothesis that commuters adapt to congestion dynamics by departing earlier or later in the morning if they would face a long congestion-related delay during the peak hour and found statistically significant evidence for it. Our estimated scheduling utility for commuters indicates that the unit cost of schedule delays is about half of the unit cost of travel time.

Perhaps the most important discovery is the large difference between travel-time profiles and indifference curves, from which we offer new insights on trip scheduling patterns and on the distribution of ideal arrival times. This finding raises some fundamental questions that cannot be easily answered in the classical bottleneck models. This paper, therefore, highlights the importance of considering the heterogeneity of location and routes when modeling congestion to solve these problems.

One limiting assumption of our model is the exogenous nature of ideal arrival times, so a possible extension would be to endogenize the formulation of ideal arrival times. For example, we could distinguish the short- and the long-run ideal arrival times, with the latter determined by the long-run preference that shapes commuters' routines in the morning periods and determines the short-run ideal arrival time (see [Peer et al., 2015](#); [Verhoef, 2020](#), for more on this idea). We also need more research investigating the topic in the same direction as our paper, using big data and new empirical tools to better apply economic models of traffic congestion and provide further insights into travel behavior.

## Bibliography

- Akbar, P., G. Duranton, V. Couture, and A. Storeygard (2019). Mobility and congestion in urban India. Unpublished paper.
- Akbar, P. A. and G. Duranton (2017). Measuring congestion in a highly congested city: Bogota. Unpublished paper.
- Anas, A. (2015). Why are urban travel times so stable? *Journal of Regional Science* 55(2), 230–261.
- Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Arnott, R., A. de Palma, and R. Lindsey (1990). Economics of a bottleneck. *Journal of Urban Economics* 27(1), 111–130.
- Arnott, R., A. de Palma, and R. Lindsey (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review* 83(1), 161–179.
- Arnott, R., A. de Palma, and R. Lindsey (1994). The welfare effects of congestion tolls with heterogeneous commuters. *Journal of Transport Economics and Policy* 28(2), 139–161.
- Bento, A., J. D. Hall, and K. Heilmann (2021). Estimating the negative externalities of traffic congestion using big data. Unpublished paper.
- Couture, V., G. Duranton, and M. Turner (2018). Speed. *Review of Economics and Statistics* 100(4), 725–739.
- Duranton, G. and A. J. Venable (2018). Place-based policies for development. World Bank Policy Research Working Paper.
- Fosgerau, M. and A. de Palma (2012). Congestion in a city with a central bottleneck. *Journal of Urban Economics* 71(3), 269–277.
- Fosgerau, M. and J. Kim (2019). Commuting and land use in a city with bottlenecks: Theory and evidence. *Regional Science and Urban Economics* 77, 182–204.

- Fosgerau, M., J. Kim, and A. Ranjan (2018). Vickrey meets Alonso: Commute scheduling and congestion in a monocentric city. *Journal of Urban Economics* 105, 40–53.
- Glaeser, E. L., S. D. Kominers, M. Luca, and N. Naik (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* 56(1), 114–137.
- Hall, J. D. (2021a). Can tolling help everyone? Estimating the aggregate and distributional consequences of congestion pricing. *Journal of the European Economic Association* 19(1), 441–474.
- Hall, J. D. (2021b). Inframarginal travelers and transportation policy. Unpublished paper.
- Hjorth, K., M. Börjesson, L. Engelson, and M. Fosgerau (2015). Estimating exponential scheduling preferences. *Transportation Research Part B: Methodological* 81(1), 230–251.
- Hörcher, D., D. J. Graham, and R. J. Anderson (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological* 95, 105–125.
- Kim, J. (2019). Estimating the social cost of congestion using the bottleneck model. *Economics of Transportation* 19, 100119.
- Kreindler, G. E. (2020). Peak-hour road congestion pricing: Experimental evidence and equilibrium implications. Unpublished paper.
- Kreindler, G. E. and Y. Miyauchi (2020). Measuring commuting and economic activity inside cities with cell phone records. Unpublished paper.
- Peer, S., J. Knockaert, and E. T. Verhoef (2016). Train commuters’ scheduling preferences: Evidence from a large-scale peak avoidance experiment. *Transportation Research Part B: Methodological* 83, 314–333.
- Peer, S., E. T. Verhoef, J. Knockaert, P. Koster, and Y. Y. Tseng (2015). Long-run versus short-run perspectives on consumer scheduling: Evidence from a revealed-preference experiment among peak-hour road commuters. *International Economic Review* 56(1), 303–323.
- Pigou, A. (1920). *The Economics of Welfare*. London: Macmillan and Co.

- Russo, A., M. W. Adler, F. Liberini, and J. N. van Ommeren (2021). Welfare losses of road congestion: Evidence from Rome. *Regional Science and Urban Economics* 89, 103692.
- Selod, S. (2021). Big Data in Transportation. An Economics Perspective. Unpublished paper.
- Small, K. A. (1982). The scheduling of consumer activities: Work trips. *The American Economic Review* 72(3), 467–479.
- Small, K. A. (2012). Valuation of travel time. *Economics of Transportation* 1(1-2), 2–14.
- Small, K. A. and E. T. Verhoef (2007). *The Economics of Urban Transportation*. Routledge, London.
- Small, K. A., C. Winston, and J. Yan (2005). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* 73(4), 1367–1382.
- Tang, C. K. (2021). The cost of traffic: Evidence from the London congestion charge. *Journal of Urban Economics* 121, 103302.
- Tarduno, M. (2021). The congestion costs of Uber and Lyft. *Journal of Urban Economics* 122, 103318.
- Verhoef, E. T. (2020). Optimal congestion pricing with diverging long-run and short-run scheduling preferences. *Transportation Research Part B* 134, 191–209.
- Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review* 59(2), 251–260.
- Walters, A. A. (1961). The theory and measurement of private and social cost of highway congestion. *Econometrica* 29(4), 676–699.
- Wardman, M. (2001). A review of British evidence on time and service quality valuations. *Transportation Research Part E: Logistics and Transportation Review* 37(2-3), 107–128.
- Yang, J., A. O. Purevjav, and S. Li (2020). The marginal cost of traffic congestion and road pricing: Evidence from a natural experiment in Beijing. *American Economic Journal: Economic Policy* 12(1), 418–453.

Table 1: Summary Statistics

	Obs	Mean (Std Dev)	Median	Min	Max
<b>Key variables</b>					
Arrival time (in minutes from midnight) <sup>a</sup>	14,544	478.19 (77.54)	478	300	675
Travel time in minutes (CHTS)	14,544	25.07 (18.50)	20	1	295
Trip distance in miles (CHTS)	14,377	12.28 (12.90)	8.38	0	389
Travel time in minutes (Google Maps) <sup>b</sup>	14,544	26.08 (17.35)	21.5	1	137
Trip distance in miles (Google Maps) <sup>c</sup>	14,544	15.30 (12.35)	11.78	0.3	81
Speed (miles per hour, Google Maps)	14,544	33.42 (10.45)	33.38	6.4	67
<b>Quantities measuring congestion</b>					
Congestion-free travel time in minutes <sup>d</sup>	14,544	21.46 (13.90)	18	0.7	81
Congestion delay, realized at chosen arrival timing (in minutes)	14,544	4.65 (6.75)	2.08	0	71
Congestion delay per trip length (minute per mile) <sup>c</sup>	14,544	0.35 (0.40)	0.2	0	4.88
Congestion delay conditional on arrival is at the peak ( $\widehat{Q^{peak}}$ )	14,544	8.36 (10.09)	4.3	0	78
<b>Selected control variables</b>					
Household income <sup>e</sup>	13,460	106.1 (68.74)	87.5	5	300
Race is white	14,544	0.73		0	1
Gender is female	14,544	0.48		0	1

Notes: a. The arrival time variable follows the decimal system and is in minutes from midnight. For example, the average arrival of 478 means 7:58 AM. b. Each commuter's Google Maps predicted travel time is the average of query outcomes between January 6 and March 18, 2020, for arrivals that fall in the interval chosen by the commuter. c. Note that trip distance may differ for given commute route (zip code pair), as Google Maps may suggest different routes. For the trip length, we use the distance of the route suggested by Google Maps at the commuter's chosen arrival time interval. d. Congestion-free travel time for each commute route (zip code pair) is estimated by taking the average of travel time outcomes for arrivals before 6:22 queried between March 19 and June 30 in 2020. e. The original income variable reported in the CHTS is categorical, but we convert it to a continuous variable by choosing the mid-point of each interval.

Table 2: Percentiles of Estimated Congestion Delays

	Delay, realized	Expected delay at the peak
5%	0.00	0.38
10%	0.17	0.82
25%	0.76	1.91
50%	2.08	4.30
75%	5.38	10.77
90%	12.81	22.42
95%	19.16	30.77
99%	32.21	46.38

*Notes:* The second column shows the distribution of the estimated individuals' realized delays (denoted by  $\widehat{Q}_i(t_i)$ ). The third column shows the distribution of the maximum queuing times that each commuter would experience if choosing the peak time on her route (denoted by  $\widehat{Q}_i^{peak}$ ).

Table 3: Testing the Adaptation Hypothesis

	(1)	(2)	(3)	(4)	(5) <sup>a</sup>	(6) <sup>a</sup>	(7)
	OLS	OLS	OLS	OLS	IV	IV	Probit
Dependent variable: Dummy for arrival in 7:52–8:52							
$\widehat{Q}^{peak}$	-0.00173*** (-4.98)	-0.000808** (-2.01)	-0.00143*** (-3.39)	-0.00265*** (-5.30)	-0.00258*** (-3.02)	-0.00241*** (-2.70)	-0.00838*** (-5.03)
Distance between home and work in miles		-0.00153*** (-4.50)	-0.00106*** (-3.00)	-0.000267 (-0.69)	-0.000297 (-0.61)	-0.000367 (-0.73)	-0.000878 (-0.71)
Number of household members			-0.00774** (-2.08)	-0.000699 (-0.18)	-0.000701 (-0.18)	-0.000704 (-0.18)	-0.00258 (-0.21)
Multi-worker household			-0.00481 (-0.54)	-0.00370 (-0.41)	-0.00369 (-0.41)	-0.00367 (-0.41)	-0.0119 (-0.43)
Has students			0.0411*** (4.03)	0.0419*** (4.04)	0.0419*** (4.06)	0.0419*** (4.06)	0.125*** (3.96)
Homeowner			0.00385 (0.37)	0.00101 (0.09)	0.00102 (0.10)	0.00103 (0.10)	0.00371 (0.11)
Annual family income in \$1,000			0.000271*** (4.03)	0.0000815 (1.13)	0.0000813 (1.13)	0.0000809 (1.12)	0.000227 (1.08)
Female			0.0717*** (9.03)	0.0434*** (4.88)	0.0434*** (4.90)	0.0435*** (4.91)	0.131*** (4.89)
College or higher degree			0.0819*** (9.67)	0.0610*** (6.36)	0.0610*** (6.38)	0.0609*** (6.38)	0.187*** (6.49)
White			0.0179** (2.00)	0.0155* (1.67)	0.0155* (1.68)	0.0156* (1.70)	0.0488* (1.69)
Has less than 5 full work days			-0.0382*** (-3.56)	-0.0387*** (-3.49)	-0.0387*** (-3.51)	-0.0387*** (-3.51)	-0.119*** (-3.43)
Flexible worker			-0.000270 (-0.02)	-0.0171 (-1.39)	-0.0171 (-1.39)	-0.0170 (-1.39)	-0.0471 (-1.26)
County-fixed effects	No	No	No	Yes	Yes	Yes	Yes
Industry dummies <sup>b</sup>	No	No	No	Yes	Yes	Yes	Yes
Occupation dummies <sup>b</sup>	No	No	No	Yes	Yes	Yes	Yes
<i>N</i>	14,544	14,544	13,460	13,133	13,133	13,133	13,132
<i>R</i> <sup>2</sup>	0.001	0.003	0.024	0.050	0.050	0.050	

Notes: Heteroskedasticity-robust t statistics are in parentheses. a. In column (5), the mean of congestion delays across commuters whose work zip code is the same as the commuter as the instrument. In column (6), the instrument is the mean of congestion delay per mile across commuter whose work zip code is the same as the commuter. b. The CHTS classifies 21 industry categories and 25 occupation categories. \* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

Table 4: Testing the Adaptation Hypothesis, Alternative Specifications

	(1)	(2)	(3)	(4)	(5)
	Early vs. Peak	Late vs. peak			
Dependent variable: Dummy for arrival in 7:52–8:52					
$\widehat{Q}^{peak}$	-0.00353*** (-5.73)	-0.00288*** (-3.49)			
$\widehat{Q}^{peak}$ per mile			-0.0252*** (-2.76)		
Log of $\widehat{Q}^{peak}$				-0.0136*** (-2.89)	
Log of $\widehat{Q}^{peak}$ per mile					-0.0142*** (-2.75)
Distance between home and work in miles	-0.000934** (-2.05)	0.00101 (1.61)	-0.00151*** (-4.67)	-0.000785** (-1.98)	-0.00164*** (-4.96)
Household and personal characteristics <sup>a</sup>	Yes	Yes	Yes	Yes	Yes
County fixed effects	Yes	Yes	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes	Yes	Yes
Occupation fixed effects	Yes	Yes	Yes	Yes	Yes
$N$	10,028	6,937	13,133	12,954	12,954
$R^2$	0.112	0.088	0.049	0.049	0.049

Notes: Heteroskedasticity-robust t statistics are in parentheses. OLS is used in all specifications. a. See Table 3 for the list of control variables. \* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

Table 5: Scheduling Utility Estimates

	(1)	(2)	(3)	(4)	(5)
$T$	-0.121*** (0.00595)	-0.131*** (0.00659)	-0.114*** (0.00915)	-0.115*** (0.00799)	-0.119*** (0.00681)
$SDE$	-0.0622*** (0.000692)	-0.0187*** (0.00230)	-0.0636*** (0.00102)	-0.0598*** (0.000975)	-0.0647*** (0.000822)
$SDL$	-0.0660*** (0.000763)	-0.0328*** (0.00237)	-0.0616*** (0.000956)	-0.0618*** (0.00101)	-0.0674*** (0.000883)
$SDE^2$		-0.000881*** (0.0000488)			
$SDL^2$		-0.000605*** (0.0000486)			
$T \times RICH$			-0.0124 (0.0121)		
$SDE \times RICH$			0.00229* (0.00139)		
$SDL \times RICH$			-0.0116*** (0.00156)		
$T \times FEMALE$				-0.0128 (0.0119)	
$SDE \times FEMALE$				-0.00494*** (0.00138)	
$SDL \times FEMALE$				-0.00891*** (0.00153)	
$T \times FLEX$					-0.0108 (0.0140)
$SDE \times FLEX$					0.00920*** (0.00149)
$SDL \times FLEX$					0.00545*** (0.00177)
$DUM15$	-0.278*** (0.0200)	-0.277*** (0.0199)	-0.278*** (0.0200)	-0.278*** (0.0200)	-0.278*** (0.0200)
Log likelihood	-17779.4	-17537.3	-17746.8	-17764.7	-17760.9
$\beta/\alpha$	0.514		0.558 (Poor) 0.485 (Rich)	0.520 (Male) 0.507 (Female)	0.544 (Nonflex) 0.428 (Flex)
$\gamma/\alpha$	0.545		0.540 (Poor) 0.579 (Rich)	0.537 (Male) 0.553 (Female)	0.566 (Nonflex) 0.477 (Flex)

*Notes:* The multinomial logit is used for estimation. The choice set includes the intervals of  $m = 2, \dots, 15$ , which covers the times from 6:22 to 9:52 in the morning. In each model, the number of cases (commuters) is 10,487, with the total number of observations in the estimation sample being 146,818 ( $= 14 \times 10,487$ ). The parameter that corrects the naive machine-learning estimates for ideal arrival times (i.e.,  $\pi$ ) is set at 0.4. Heteroskedasticity-robust standard errors are in parentheses. \* $p < 0.10$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

Table 6: Effects of  $\pi$  on the marginal rate of substitution

$\pi$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\beta/\alpha$	1.46	0.85	0.62	0.51	0.47	0.45	0.47	0.53	0.69	1.43
$\gamma/\alpha$	1.58	0.91	0.66	0.55	0.50	0.49	0.52	0.61	0.85	1.91

*Notes:* The linear specification of the scheduling utility (the model in column (1) in Table 5) is estimated at each  $\pi$  value. The number of cases (commute choices) in the estimation sample is 10,487, with the total number of observations being 146,818 ( $= 14 \times 10,487$ ).

Figure 1: Trip Scheduling of the Commuter

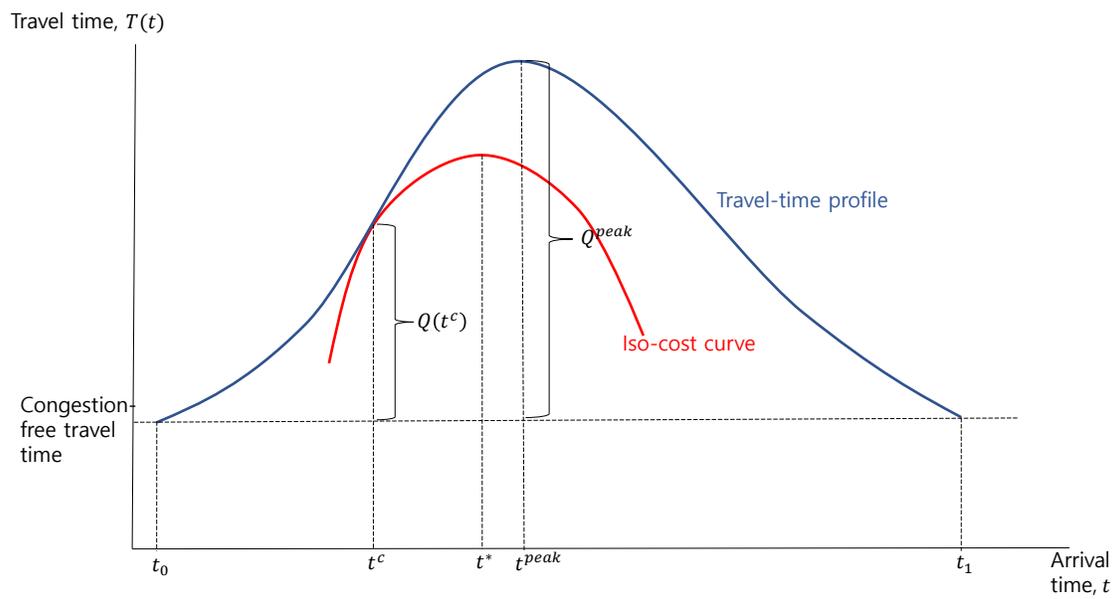
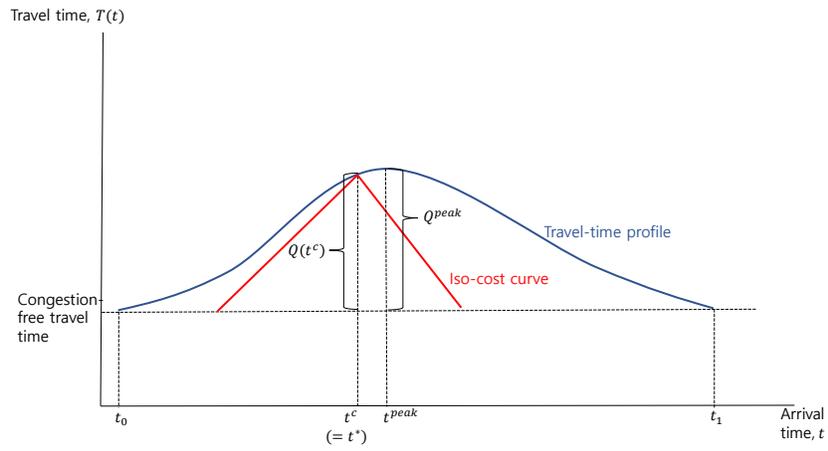


Figure 2: The Effect of Travel-Time Profile on Scheduling Choice

(a) Low congestion level



(b) High congestion level

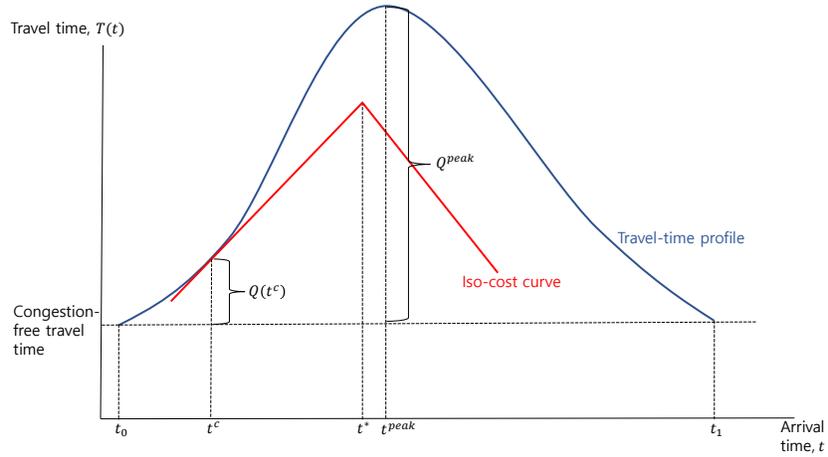
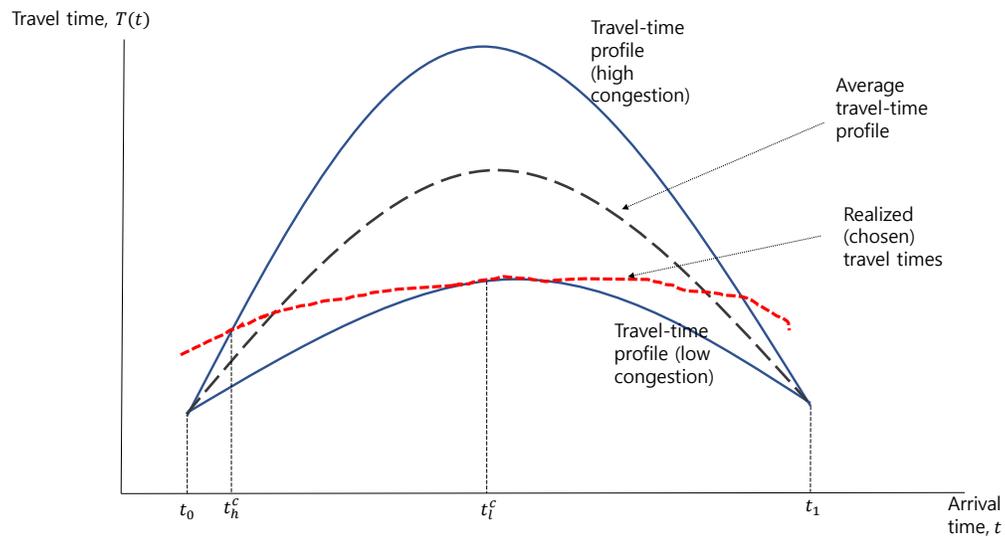
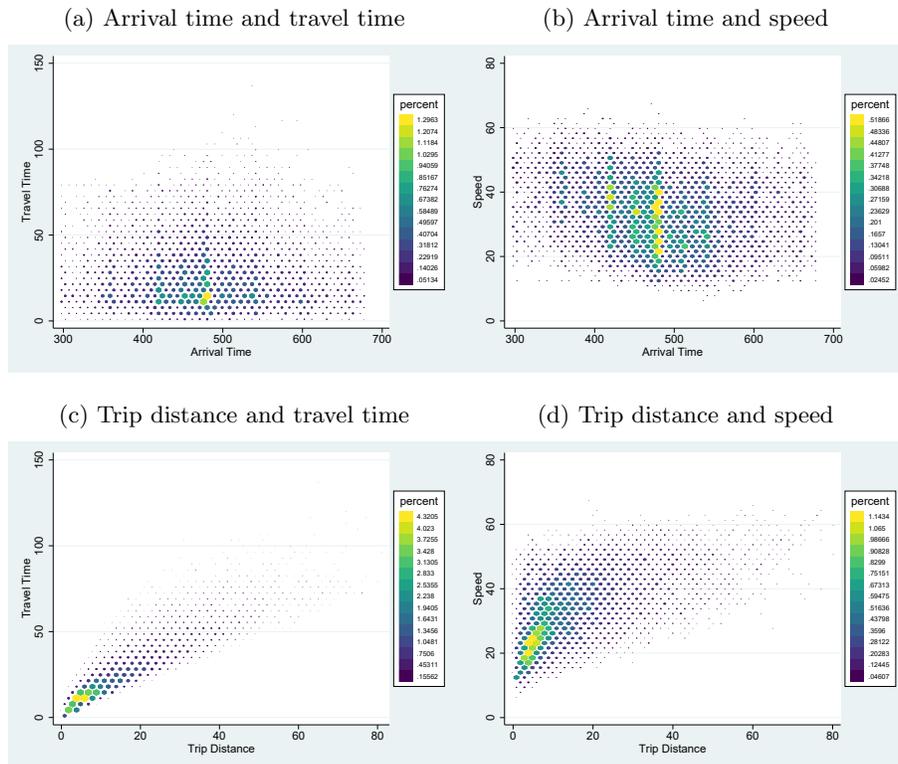


Figure 3: Exemplified Average and Realized Travel-Time Profiles



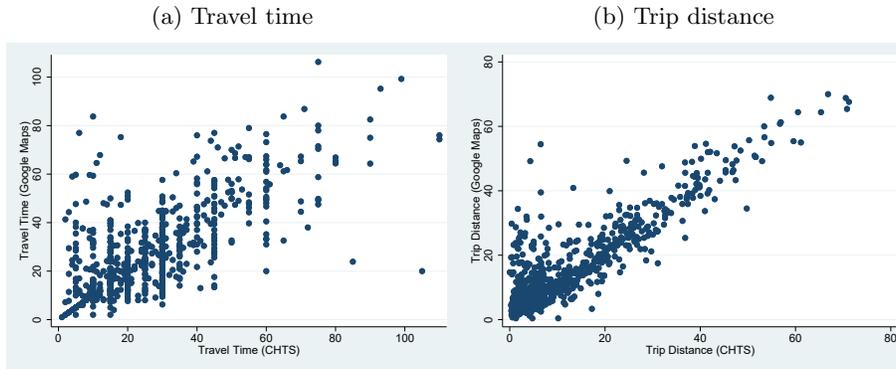
*Notes:* In this figure, we exemplify the average travel-time profile faced by commuters (long-dashed curve) and scatter plots for arrival time and realized travel times (short-dashed curve). The figure illustrates two different travel-time profiles (high and low congestion) and corresponding arrival time choices ( $t_h^c$  and  $t_l^c$ , respectively). There are two points on the short-dashed curve corresponding to these individuals' choices. Note that while we visualize only two travel-time profiles in this figure, the full short-dashed curve is drawn based on the assumption that there is a continuum of different individuals facing their own travel-time profiles. As a result, as depicted in the figure, the short-dashed curve may be outside of the range between the two travel-time profiles.

Figure 4: Scatter Diagrams Between Variables



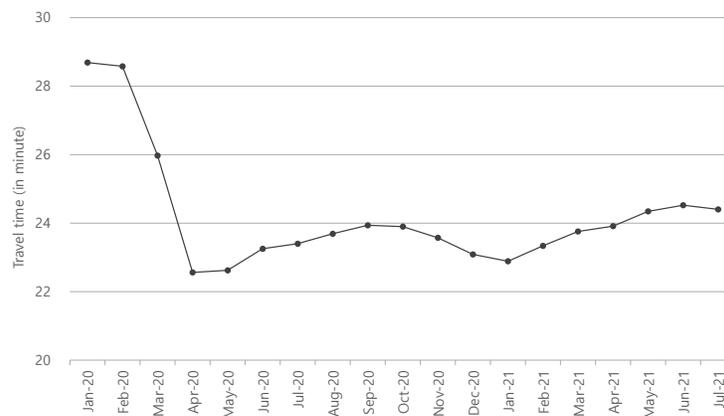
*Notes:* The scatter diagrams are drawn to visualize the relationships between different variables. Travel times and trip distances predicted at the commuters' chosen arrival time intervals by Google Maps are used to draw scatter diagrams, while the arrival time variable is from the CHTS. The CHTS reported travel time and trip distance exhibit similar data patterns. The speed variable plotted in panels (b) and (d) are calculated by dividing the trip distance (measured by Google Maps) by travel time (predicted by Google Maps) and multiplying it by 60 to normalize into a per-hour scale.

Figure 5: Degree of Fit of Google Maps to CHTS Data



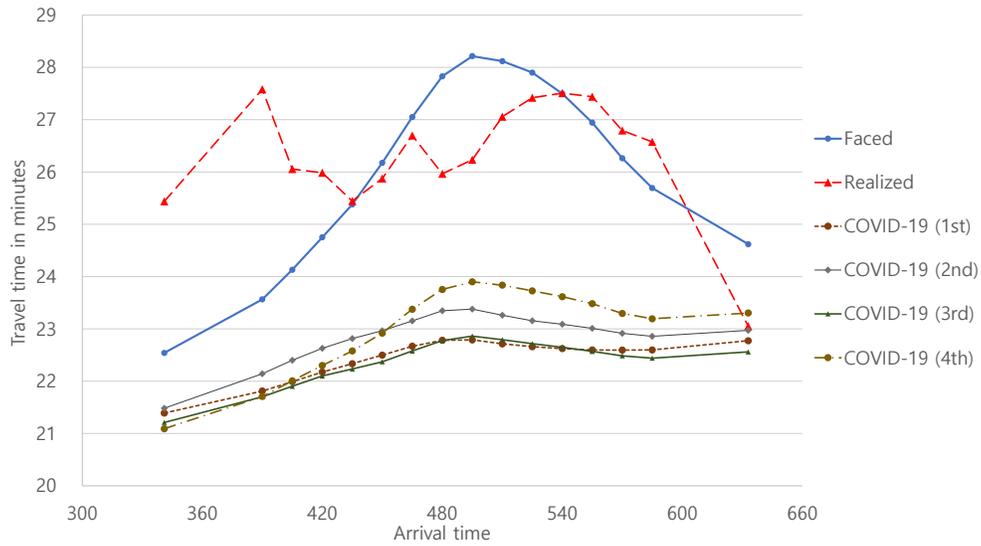
*Notes:* The scatter diagrams are drawn to evaluate the degree of fit of the Google Maps predicted values to the CHTS reported values of the key variables. Only 5% of observations from the entire sample are used to draw the scatter diagrams.

Figure 6: Average of Expected Travel Time in the Peak Hour by Month



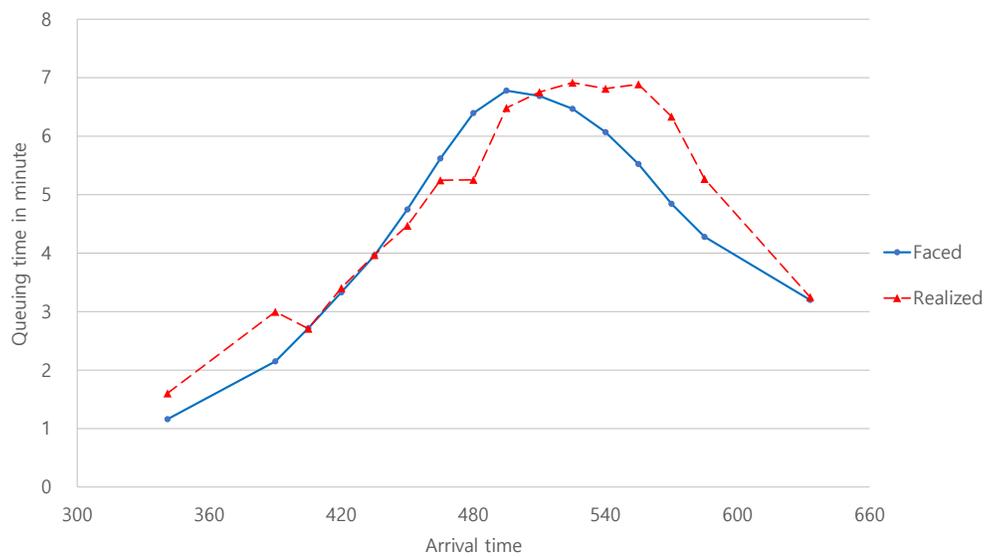
*Notes:* Each point plots the monthly average of expected travel times at the peak hour of the commuters in the sample. The expected travel time of the peak hour (7:52 AM–8:52 AM) of each commuter is first calculated by taking the mean of Google Maps travel times using trips whose arrival times are in the peak hour over the month, which is then averaged out across commuters to give each point.

Figure 7: Travel-Time Profiles Averaged Out Across Commuters



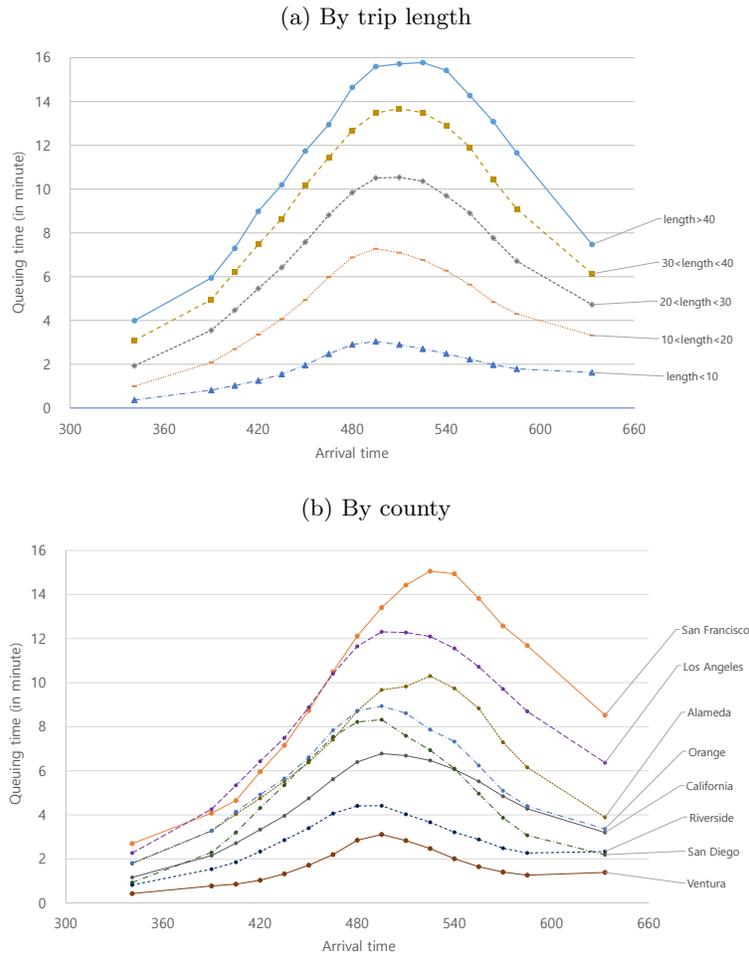
*Notes:* The arrival time variable represented on the horizontal axis is in minutes, and time 00:00 is normalized at 0. The solid curve (blue colored) illustrates the average of travel-time profiles of the commuters in the sample. For this, we first construct each person’s travel-time profile (systematic travel-time prediction conditional on the commuter’s arrival time being in each interval) by taking the mean of travel times predicted by Google Maps using counterfactual trips whose arrival times fall in each interval over the data collection period. These are then averaged out across commuters by arrival time interval to give each point in the solid curve. The data collection period from Google Maps covers January 6–March 18, 2020. Each point on the long-dashed (red colored) curve represents the average of the travel-time values calculated only using the commuters who chose that interval. The average travel-time profiles faced by travelers are also drawn in the figure by selecting different data collection periods to illustrate how the travel-time profiles have changed since the COVID-19 measures entered into force in California: COVID-19 1st period covers March 19–July 31, 2020; COVID-19 2nd period covers August 1–November 30, 2020; COVID-19 3rd period covers December 1–March 31, 2021; and COVID-19 4th period covers April 1, 2021–July 28, 2021.

Figure 8: Average Congestion-Delay Profiles Faced and Realized



*Notes:* The average congestion-delay profile faced by commuters is illustrated by the solid blue curve. Each commuter's expected congestion-delay profile is first constructed by subtracting the estimated congestion-free travel time from the Google Maps constructed travel-time prediction of each arrival time interval. These individuals' expected queuing times at respective arrival time intervals are then averaged out across commuters in the sample to construct the solid blue curve. Each point on the red dotted curve indicates the average of the expected queuing times calculated from only commuters who chose that arrival time interval.

Figure 9: Variation of Congestion



*Notes:* Panel (a) shows how congestion-delay profiles (faced by individuals) differ by trip length. We classify commuters into five different groups sorted by trip length (the length of each pair of zip codes is the average of distances of different routes suggested by Google Maps by arrival time interval for the commuter) and draw the average congestion-delay profiles of commuters in different groups. Panel (b) draws the congestion-delay profiles for different counties. We group trips by the destination county. Of the 57 counties in California, several large counties are selected in this figure. See Table A1 for the number of observations used to draw each profile.

Figure 10: Difference in the Distribution of Arrival Times

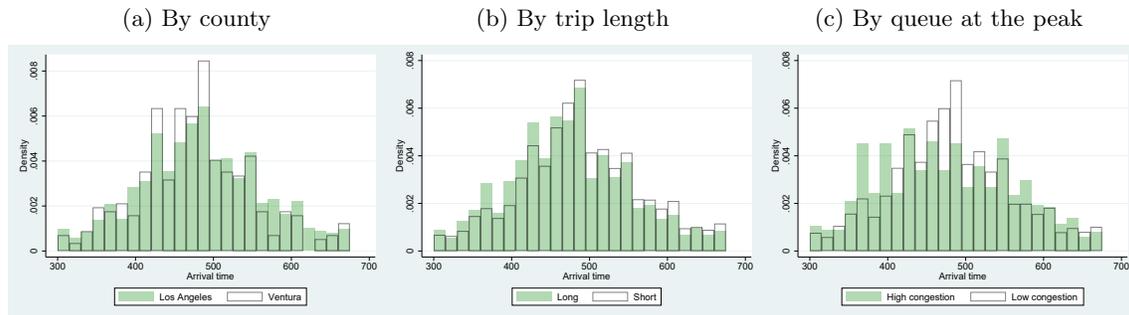
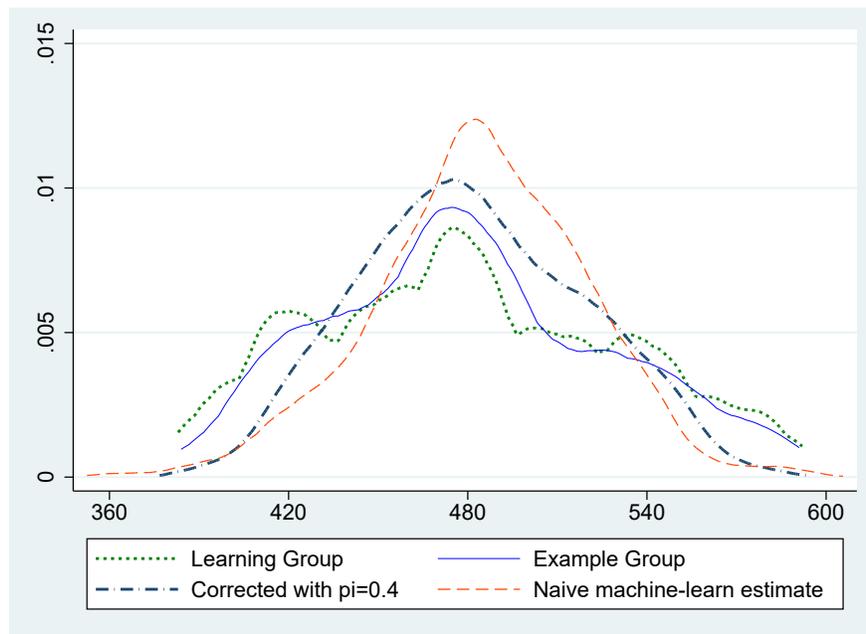
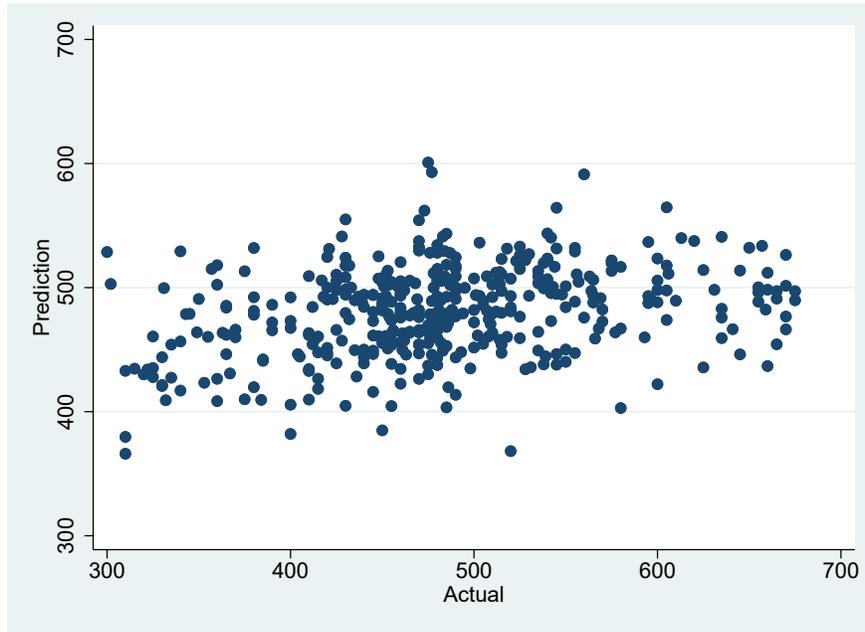


Figure 11: Distribution of Arrival Times



*Notes:* This figure shows the effect of  $\pi$  (weight placed on the machine-learning estimates for ideal arrival times) on the distribution of the assigned ideal arrival times. As  $\pi$  is smaller, because the actual arrival times are more dispersed, the distribution becomes more dispersed as well. We can see that with  $\pi$  being around 0.3, the distribution of finally assigned ideal arrival times are closest to the distribution of arrival times of the example group (who are assumed to arrive at their ideal arrival times).

Figure 12: Testing Prediction Power



*Notes:* In this figure, we test the prediction power. To do so, we randomly select 30% of the 1,388 observations in the example group and predict their arrivals using the rest of the group (70%). Because the selected group's arrival times are known as well as predicted values, we can see the fitness. We plot scatter plots for the actual arrival times (that are assumed to be ideal for this group) and the predicted values.

Figure 13: Conjecture

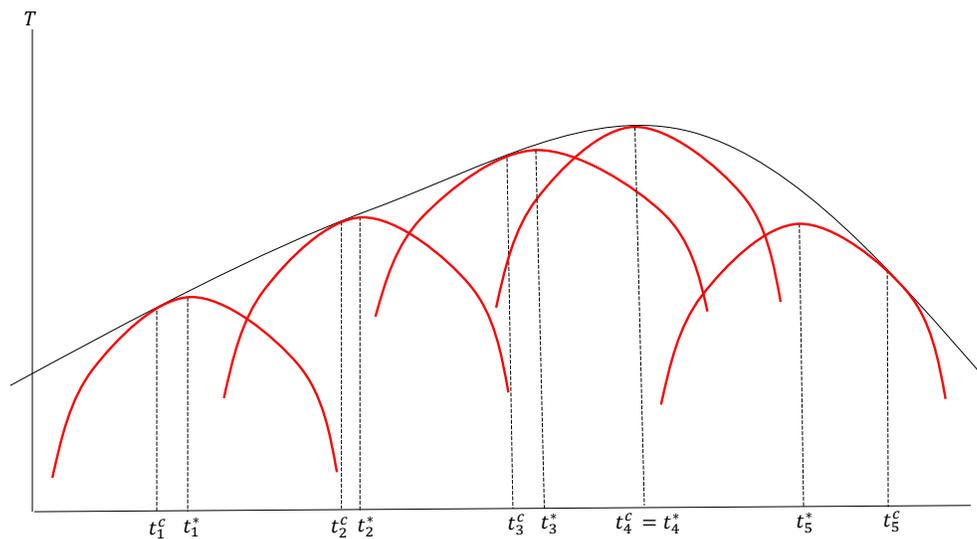
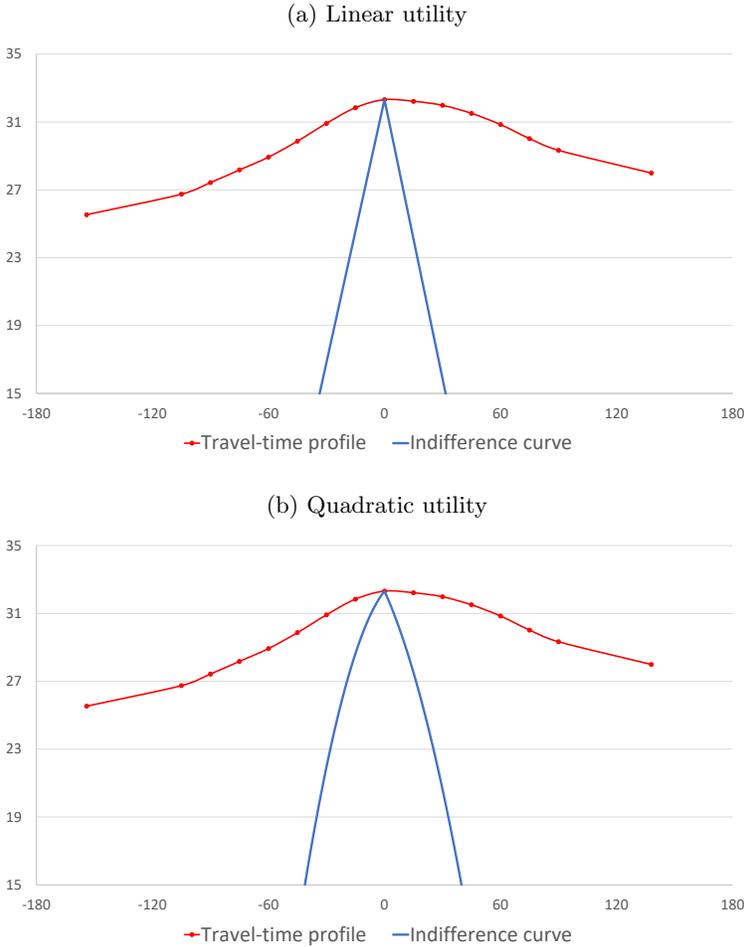


Figure 14: Scheduling Utility and Travel-Time Profiles



# Appendices

## A. Proof of Proposition 1

**Proof.**

1. Points above the travel-time profile are not optimal because the commuter could drive faster at any chosen trip timing and reduce commuting cost. Points below the travel-time profile are not feasible. Given the larger curvature of the iso-cost curves and  $t_0 < t^* < t_1$ , there is at least one point at which an iso-cost curve and the travel-time profile meet. Therefore, it is sufficient to examine only the points on the travel-time profile. Given the curvature difference, the iso-cost curve is steeper than the travel-time profile at  $t_0$ , under which the commuter could have a lower-positioned iso-cost curve by arriving later than  $t_0$ , so  $t_0$  is not optimal. The same reasoning is applied to show  $t_1$  is not optimal. As another possibility, if an iso-cost curve crosses the travel-time profile at two points between  $t_0$  and  $t_1$ , then neither can be optimal because the commuter could reduce the cost by choosing a time closer to  $t^*$ . The remaining possibility is that an iso-cost curve and the travel-time profile meet at a unique point, at which these two curves are tangent.
2. We first rule out the possibility that the iso-cost curve is above the travel-time profile, as it is not optimal. Since the travel-time profile is more convex, the iso-cost curve and the travel-time profile always cross twice as long as the crossing points are not among  $t_0$ ,  $t_1$ , or  $t^*$ . If the two curves cross at any two points, then the commuter could shift the iso-cost curve down by choosing a time closer to  $t_0$ ,  $t_1$ , or  $t^*$ . The commuter does so until the iso-cost curve touches  $t_0$  or  $t_1$ , or  $t^*$ , which implies that the only possible optimal choices are  $t_0$  or  $t_1$  or  $t^*$ .

■

## B. Further Notes on Data Construction

The first point pertains to how we deal with commuters traveling within the same zip code because Google Maps would be unable to differentiate the locations between origin and destination for these commuters (with their expected travel time being always zero, since Google Maps would have the same zip code for home and work). Among the 9,127 different zip code pairs in the sample, 668 pairs have the same work and home zip code. For their home zip codes, we match the nearest home zip codes among other commuters traveling to the same work zip code as the commuter. Because the matched zip code is different and thus may overestimate the trip distance, we rescale the travel time outcomes using the ratio of the CHTS reported travel time and the Google Maps travel time predicted at the worker's chosen trip timing.

The second point pertains to the data collection algorithm we use to work with Google Maps data. We first set up the list of zip code pairs, and we then queried travel times for each of them at random timings between 5:00 AM and 11:00 AM on each weekday from January 6, 2020 to July 28, 2021. We program random selections of a zip code pair queried by Google Maps because we want to avoid the bias that may arise due to a systematically higher frequency of particular routes chosen at particular (peak) times within travel days; we try to collect full travel-time profiles for as many routes as possible.

Third, we describe in detail how we construct systematic travel time expectations of individuals by alternate arrival time interval. The notion in our framework is that each commuter would respond to her own systematic anticipation of travel times at her alternate choices, which would be formed from her own experiences of traveling over many days. Our goal is to estimate this anticipated menu of travel times, not one on a specific day, because it would involve unexpected delays due to a weather shock or accidents. This is why we collected the data over a long horizon. However, out of the entire data collection period from January 6, 2020, to July 2021, we had to exclude the data collected subsequent to the lockdown policy enacted in California on March 19, 2020, in response to the soaring COVID-19 cases. Figure 6 shows a significant drop in travel times around this date and the trend has not fully recovered as of the summer of 2021. It is appropriate to exclude the travel time outcomes during these periods, because the commuters in the CHTS sample were subject to travel-time profiles similar to those of the pre-COVID-19 period and we want to explore scheduling choices under the congested situation. Thus, we limit our usage of Google Maps data to the “normal” period before March 19, 2020. We are left with 1,255,426 queried trips from January 6 to March 18, 2020, and we use these to construct the travel-time profiles during the congested (normal) time periods. When designing this research, we had not anticipated COVID-19, but we were fortunate to have begun data collection before the pandemic, as we obtained enough travel time predictions for each pair of home and work zip codes to construct the systematic travel-time profile for each commuter. Queried trips subsequent to COVID-19 lockdown measures were instead used to construct the congestion-free travel time for each route.

Finally, note that our final sample consists of 14,544 commute observations that have a “full” travel-time profile with no missing values on travel times in any interval or congestion-free travel period. This number is fewer than the 16,376 commuters in the raw CHTS sample: we lose 1,832 observations. The removed observations may have a missing value for the mean travel time in an interval for the route, because our query is random at each timing, so these zip code pairs failed to fill all the defined intervals. We drop the travel time menu for a route if it misses at least one value in the intervals to ensure that commuters in our sample face the “full” menu of travel times. Note however that the selection is effectively random, given the data collection algorithm.

## C. Additional Table

Table A1: Ranking of Congestion Level by County

County	Population size <sup>a</sup>	Obs (CHTS)	Mean of $\widehat{Q}_i^{peak}$	Mean of per-mile $\widehat{Q}_i^{peakb}$
San Francisco	881,549	322	17.52	1.59
Los Angeles	10,039,107	2839	14.64	1.08
Santa Clara	1,927,852	1159	13.02	0.91
Alameda	1,671,329	664	12.52	0.82
San Mateo	766,573	471	12.20	0.80
Marin	258,826	140	10.31	0.72
Orange	3,175,692	962	10.25	0.73
San Diego	3,338,330	598	9.70	0.68
Sacramento	1,552,058	359	8.74	0.65
Contra Costa	1,153,526	461	7.92	0.59
San Bernadino	2,180,085	547	6.25	0.41
Santa Cruz	273,213	188	5.65	0.53
Riverside	2,470,546	441	5.35	0.39
Monterey	434,061	365	4.88	0.50
Sonoma	494,336	305	4.36	0.38
Napa	137,744	130	4.34	0.28
Solano	447,643	226	3.74	0.25
Ventura	846,006	378	3.56	0.33
Lassen	30,573	34	3.47	0.18
Stanislaus	550,660	166	3.38	0.35
Fresno	999,101	440	3.31	0.33
Yuba	78,668	61	3.20	0.18
San Joaquin	762,148	211	3.14	0.23
Yolo	220,500	86	2.94	0.20
Placer	398,329	151	2.94	0.31
Santa Barbara	446,499	218	2.86	0.24
Kern	900,202	369	2.56	0.18
El Dorado	192,843	98	2.56	0.27
Butte	219,186	122	2.51	0.23
Humboldt	135,558	95	2.36	0.28
Merced	277,680	144	2.32	0.18
San Benito	62,808	41	2.24	0.19
Madera	157,327	78	2.17	0.16
Sutter	96,971	53	2.05	0.12
Shasta	180,080	49	1.94	0.16
Tuolumne	54,478	62	1.90	0.17
Sierra	3,005	7	1.88	0.20
Siskiyou	43,539	46	1.87	0.12
Amador	39,752	32	1.80	0.13
Nevada	99,755	51	1.79	0.14
Tulare	466,195	270	1.53	0.15
San Luis Obispo	283,111	276	1.49	0.11
Lake	64,386	44	1.41	0.12
Calaveras	45,905	29	1.37	0.08
Kings	152,940	85	1.17	0.06
Mendocino	86,749	56	1.09	0.06
Mariposa	17,203	25	1.03	0.06
Glenn	28,393	44	0.96	0.13
Tehama	65,084	40	0.92	0.08
Imperial	181,215	205	0.87	0.07
Mono	14,444	29	0.80	0.04
Colusa	21,547	30	0.72	0.04
Plumas	18,807	47	0.56	0.04
Modoc	8,841	26	0.52	0.04
Trinity	12,285	27	0.50	0.06
Del Norte	27,812	74	0.20	0.02
Inyo	18,039	68	0.08	0.00
All CA	39,512,223	14544	8.36	0.62

*Notes:* This table shows the levels of congestion by work site county. Commuters are sorted by the county where their workplaces are located, and the mean of peak queuing times and the mean of per-mile peak queuing time by the county groups are calculated. While the table sorts the counties (the mean of peak queuing times) from the highest to the lowest, note that the ranking based on the mean of per-mile queuing times is almost the same. a. The population size is of 2019 (Source for County Population Totals: 2010–2019, US Census Bureau). b. In calculating per-mile  $\widehat{Q}_i^{peak}$ , we use the Google Maps distance for the suggested route at each commuter's chosen arrival time.