ABSTRACT. We need structural models of traffic congestion to answer a wide variety of questions, but the standard models fail to match the data on travel times across the day. I establish the nature and magnitude of the problem, and show its source lies in how we model traveler preferences, not in the specifics of the congestion technology. The poor fit of the models suggests that we are abstracting away from features with a first-order impact on model predictions, which limits our ability to use these models to evaluate counterfactuals quantitatively and—when travelers are heterogeneous—qualitatively as well. I explore several ways of improving the fit of these models, concluding with recommendations for tractable and intuitive ways of doing so.

Keywords: Structural model, Congestion, Model fit, Calibration, Dynamic, Bottleneck model, Traffic

JEL Codes: R4, H4
1. Introduction

Arnott et al. (1993) persuasively argue it is important to use structural models of traffic congestion as the standard static model “contains ambiguities and is poorly specified” (p. 161). These structural models have been used extensively in theoretical analysis and simulations to address a variety of important questions, including optimal highway capacity (Arnott et al. 1993), the value of travel time information (Arnott et al. 1999; Khan and Amin 2018), optimal parking prices (Fosgerau and de Palma 2013), how tolls affect urban spatial structure (Gubins and Verhoef 2014; Takayama and Kuwahara 2017; Fosgerau et al. 2018), which road segments should be tolled (de Palma et al. 2004), non-price mechanisms for addressing congestion (Nie 2015), and the distributional consequences of road tolls (van den Berg and Verhoef 2011; Fosgerau and Small 2017; Hall 2018; Kreindler 2018).¹

These structural models are inherently dynamic and explain the evolution of travel times across the days. As such, these models have at their heart the decision by travelers of when to travel. Travelers choose between arriving on-time, and facing, in equilibrium, long travel times, or arriving early or late in exchange for shorter travel times. The low travel times off-peak compensate travelers for their cost of arriving early or late. This trade-off is similar to that between the price of housing (per square foot) and commuting costs in models of urban structure (such as the monocentric city model of Alonso (1964), Mills (1967), and Muth (1969)), which explain the decline in the price of housing (per square foot) with distance to the city center as a compensating differential for commuting costs.

This paper’s contribution is to, first, show that structural models of congestion fail to accurately predict travel times using reasonable values for preference parameters; second, to show that this failure affects the theoretical predictions of the model; and, third, show how to improve the models so that they do fit the data.

I start by comparing the predicted travel times from standard structural models of congestion to data, finding that while they generate a travel time profile that has the right general shape, they fit observed travel times poorly. In particular, under standard assumptions for parameter values, the predictions for either peak travel times or the length of rush hour are off by at least an order of magnitude. The poor fit suggests that these models are abstracting away from features that have a first-order impact on model predictions, which limits our ability to use them to evaluate counterfactuals.

¹Other important questions addressed include the effects of second-best pricing (van den Berg 2014; Lindsey et al. 2012), how tolls affect unemployment and the spatial distribution of activity across a region (Vandyck and Rutherford 2018), how teleworking impacts congestion (Gubins and Verhoef 2011), and the cost of travel time variability (Noland and Small 1995; Fosgerau and Engelson 2011; Engelson and Fosgerau 2016). Furthermore, the time-of-use decision at the heart of structural models of traffic congestion is also used in models of transportation by train (Kraus and Yoshida 2002; De Borger and Fosgerau 2012; de Palma et al. 2017) and airplane (Silva et al. 2014; de Palma et al. 2018; Blondiau et al. 2018).
Comparing the theoretical predictions for travel time to data highlights three specific problems. First, travel times climb and fall far slower than the models predict; second, travel times fall too slowly after the peak relative to how quickly they climb before the peak; and, third, travel times are essentially flat at the peak.

The source of the poor fit lies almost entirely with how we model traveler preferences over different arrival times, and has little to do with the model of congestion itself. Thus these problems affect the bottleneck model of Vickrey (1969) and Arnott et al. (1993), as well as other structural models such as those of Henderson (1974) and Chu (1995), the cell transmission model of Daganzo (1994), and the hydrodynamic traffic flow model of Lighthill and Whitham (1955) and Richards (1956). This is because traveler preferences determine the slope of the travel time profile, while the congestion technology then determines how many travelers are traveling at each point in time.

I explore multiple solutions to each of the three problems, coming to the following recommendations for improving the ability of structural models of congestion to fit the data. If the goal is to quantify an outcome with homogeneous travelers, then it is simplest to use significantly lower parameter values for travelers’ schedule delay costs than are typically assumed. If the goal is a quantitative or qualitative result with heterogeneous travelers, then allow for inframarginal travelers by adding the assumptions that travelers’ desired arrival times are continuously distributed and that travelers differ in the flexibility of their schedules. Furthermore, use parameter values that imply the cost of being late is less than the cost of being early. These solutions do not fix the problem of travel times being essentially flat at the peak, however, its magnitude is small after fixing the other two. If it is to be fixed, it is most tractably fixed by allowing for travelers to be indifferent between arrivals in a small window around their desired arrival time.

Of course, as Friedman (1953) argues, we do not expect models to be perfect representations of reality, but useful abstractions that help us understand reality better. The problems I identify do not affect all uses of structural congestion models. However, I show they limit our ability to use these models to quantify magnitudes of some outcome (e.g., the social welfare gains from some policy) and derive qualitative results with heterogeneous travelers.

The remainder of the paper is structured as follows. After introducing the model (Section 2), I document the three problems with the fit of the model to the data (Section 3), explore solutions to these problems (Section 4), and show these problems affect our theoretical predictions (Section 5). I conclude in Section 6.

---

2These models essentially only differ in how they model congestion. In the bottleneck model travelers only slow down those who come after them, while in Henderson (1974) and Chu (1995) travelers only slow down those traveling at the same time as them. In the cell transmission and Lighthill and Whitham (1955)-Richards (1956) models travelers slow down both those who come after and those traveling at the same time. These last two models are less tractable, and only recently have been combined with models of traveler preferences (e.g., Han et al. 2011; Ukkusuri et al. 2012; Friesz et al. 2013).
2. Model

I start with the model of traveler preferences of Vickrey (1969), which is used throughout the literature, while making minimal assumptions on the congestion technology.3

2.1. Congestion technology. There is a single road connecting where people live to where they work. Travel time for a traveler arriving at \( t \) is

\[
T(t) = T^f + T^v(t),
\]

where \( T^f \) is fixed travel time—the amount of time it takes to travel the road absent any congestion—and \( T^v(t) \) is variable travel time. I assume \( T^v(t) \) is continuous everywhere and differentiable almost everywhere.

For most of the results in this paper, this is all of the structure we need on the congestion technology. When exploring whether adding uncertainty helps fix the empirical problems with structural congestion models, I modify the model to add uncertainty; and I use the bottleneck model when discussing examples of how the problems identified in this paper affect quantitative and qualitative results.

2.2. Traveler preferences. Travelers choose when to arrive at work to minimize the cost of traveling.4 They dislike two aspects of traveling: travel time and schedule delay—that is, arriving earlier or later than desired. These costs combine to form the trip cost; the trip cost of arriving at time \( t \) for a traveler of type \( i \) is

\[
p_i(t) = \alpha_i T(t) + D_i(t^* - t)
\]

where \( \alpha \) is the cost per unit time traveling (i.e., the traveler’s value of time) and \( D_i \) is type \( i \)’s schedule delay cost function. Schedule delay costs are piecewise linear,

\[
D_i(x) = \begin{cases} 
\beta_i & x \leq 0 \\
-\gamma_i & x > 0 
\end{cases}
\]

3There are other ways of modeling traveler preferences that are similar to that above. Most prominent is the utility-theoretic models of Vickrey (1973) and Fosgerau and Engelson (2011), where traveler preferences are defined in terms of the per-minute utility of time spent traveling, at home, and at work. If the utility rate of traveling and time at home are constant, and the utility rate at work is piecewise-constant with a discontinuity at \( t^* \), then this is isomorphic to what is above. In this case \( \alpha \) is the difference between the per-minute utility at home and the per-minute utility while driving, \( \beta \) is the difference between the per-minute utility of being at home and the per-minute utility of being at work before \( t^* \), \( \gamma \) is the difference between the per-minute utility of being at home and the per-minute utility of being at work after \( t^* \).

4Nearly all papers using structural congestion models focus on the morning commute. de Palma and Lindsey (2002) point out that a key difference between the morning and evening commutes may be that in the evening travelers have a preferred departure time rather than a preferred arrival time. The first and third problems discussed below are also issues with the evening commute, and the improvements that deal with those problems are likewise relevant to the evening commute.
where $\beta$ is the cost per unit time early to work, and $\gamma$ is the cost per unit time late to work. Each of these parameters represents how much a traveler is willing to pay in money to reduce travel time or schedule delay by one unit of time.

Each individual traveler has zero mass, and there is mass $N_i$ of travelers of type $i$.

When exploring different ways of improving the fit of structural congestion models, I consider various extensions to this classic formulation of preferences.

2.3. Definition of equilibrium. The relevant equilibrium concept is that of a perfect-information, pure-strategy Nash equilibrium, in which no traveler can reduce his trip cost by changing his arrival time.

3. Identifying the problems

To better identify the problems, it is helpful to focus on three ratios: $\beta/\alpha$, $\gamma/\alpha$ and $\gamma/\beta$. I first show that these ratios have simple economic interpretations, which helps us assess reasonable values for them, and discuss the commonly assumed values for them. I then show how these ratios map into the empirical travel time profile, and use data from three cities to compare the theoretically predicted travel times given commonly assumed values for these ratios to the empirical travel time profile.

The ratios $\beta/\alpha$ and $\gamma/\alpha$ are a traveler’s willingness to pay in travel time to reduce schedule delay (early and late respectively) by one unit of time, and provide a measure of the inflexibility of his schedule. As in Hall (2018), define $\beta_i/\alpha_i$ as type $i$’s inflexibility. If a shift worker is late he generally faces penalties and when he is early he passes the time talking with co-workers. Since there is not much difference for the shift worker between spending time traveling or being at work early, his $\beta/\alpha$ is close to one. Similarly, due to the penalty when late, $\gamma/\alpha$ is large. In contrast, an academic can start working whenever she gets to the office and so has a very low marginal disutility from being early or late and so her $\beta/\alpha$ is closer to zero. Typical assumptions, as summarized in Table 1, are for mean inflexibility to be greater than one-half, and maximum values near one.

The ratio $\gamma/\beta$ represents the relative cost of being late to early. To the best of my knowledge, it is always assumed that arriving late is worse than arriving early, so that $\gamma/\beta > 1$, with typical values near four, as Table 1 shows.

Not only are the standard assumptions on $\beta/\alpha$ and $\gamma/\beta$ intuitively reasonable, they also have empirical evidence supporting them. Small (1982) estimates a discrete choice model of travelers’ choice of when to arrive at work, and finds results supporting these assumptions. Indeed, many of the papers listed in Table 1 cite Small (1982) in support of their assumptions.

5 The largest possible value of $\beta/\alpha$ is 1, as if a traveler fundamentally prefers being in their vehicle to arriving early at their destination, they can simply sit in their car in the parking lot or circle the block. Thus the cost of arriving early cannot be greater than the cost of travel time, so $\beta \leq \alpha$ and $\beta/\alpha \leq 1$. 

---

[5] The largest possible value of $\beta/\alpha$ is 1, as if a traveler fundamentally prefers being in their vehicle to arriving early at their destination, they can simply sit in their car in the parking lot or circle the block. Thus the cost of arriving early cannot be greater than the cost of travel time, so $\beta \leq \alpha$ and $\beta/\alpha \leq 1$. 

Table 1. Assumed parameter values

<table>
<thead>
<tr>
<th>Paper</th>
<th>$\beta/\alpha$</th>
<th>$\gamma/\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arnott et al. (1993)</td>
<td>0.61</td>
<td>3.9</td>
</tr>
<tr>
<td>de Palma and Lindsey (2002)</td>
<td>[0.29, 0.94]</td>
<td>4</td>
</tr>
<tr>
<td>de Palma et al. (2004)</td>
<td>0.6</td>
<td>4.17</td>
</tr>
<tr>
<td>Fosgerau and Karlström (2010)</td>
<td>0.5</td>
<td>4</td>
</tr>
<tr>
<td>Liu and Nie (2011)</td>
<td>{0.61, 0.78}</td>
<td>{3.04, 3.90}</td>
</tr>
<tr>
<td>van den Berg and Verhoef (2011)</td>
<td>[0.33, 0.99]</td>
<td>3.9</td>
</tr>
<tr>
<td>Lindsey et al. (2012)</td>
<td>0.61</td>
<td>3.9</td>
</tr>
<tr>
<td>Tian et al. (2013)</td>
<td>0.19, 0.98</td>
<td>5.25</td>
</tr>
<tr>
<td>Gubins and Verhoef (2014)</td>
<td>0.5</td>
<td>4</td>
</tr>
<tr>
<td>Xiao et al. (2014)</td>
<td>0.61</td>
<td>3.9</td>
</tr>
<tr>
<td>Takayama and Kuwahara (2017)</td>
<td>{0.3, 0.4, 0.45, 0.5}</td>
<td>4</td>
</tr>
<tr>
<td>Verhoef (2017)</td>
<td>0.5</td>
<td>4</td>
</tr>
<tr>
<td>Khan and Amin (2018)</td>
<td>0.61</td>
<td>3.9</td>
</tr>
</tbody>
</table>

These preference parameters determine the shape of the travel time profile, $T(t)$. The first-order condition for choosing arrival time, $p'_i(t) = 0$, implies

\[
T'(t) = \frac{\beta_i}{\alpha_i} \quad \text{if } t < t^*, \quad (2)
\]

\[
T'(t) = -\frac{\gamma_i}{\alpha_i} = -\frac{\gamma_i \beta_i}{\beta_i \alpha_i} \quad \text{if } t > t^*, \quad \text{and} \quad (3)
\]

\[-\frac{\gamma_i}{\alpha_i} \leq T'(t) \leq \frac{\beta_i}{\alpha_i} \quad \text{if } t = t^*. \quad (4)
\]

These impose the standard requirement that a traveler’s marginal rate of substitution be tangent to the budget line, unless at a corner solution.

For someone to be arriving at any given time $t$, one of (2)–(4) must be satisfied, and since someone is arriving at all times during rush hour, one of these equations must be satisfied at all times during rush hour. Thus the slope of the travel time profile is directly governed by travelers’ preferences, and we can learn about travelers’ preferences by looking at empirical travel time profiles.

I measure the travel time profile in three cities, Los Angeles, Chicago, and Boston, using data from a different source for each one.\(^6\) Figure 1 compares the model predicted travel times using the parameters typically assumed in the literature to travel times on California State Route 91 from the center of Corona to the junction of SR-91 and I-605.\(^7\)

\(^6\)I use data from California Department of Transportation (2014), Illinois Department of Transportation (2017), and data collected from Google Maps.

\(^7\)I choose this specific segment because it roughly represents the median commute for those living in Corona who use SR-91, as calculated using data from Sullivan (1999). I calculate travel times for all business days in 2004 using loop detector data from the California Department of Transportation’s Performance Measurement System (“PeMS”, 2014).
Figure 1. Travel times predicted by theory compared to average travel times for CA SR-91W for postmiles 16–42 for 2004. Data from Caltrans PeMS.

Notes: For the theoretical predictions I assume $\beta/\alpha = 0.6$ and $\gamma/\beta = 3.9$. The higher theoretically predicted travel times come from matching the empirical length of rush hour and the lower theoretically predicted travel times come from matching the empirical peak travel time. The reported slopes for the empirical travel time profile are the steepest slopes over an hour interval.

As Figure 1 shows, using the standard assumptions about parameter values, the model fails to match the data. Matching peak travel times requires predicting rush hour is only 46 minutes long, and matching the length of rush hour requires predicting variable travel times seven times longer than those seen in the data. While the model predicted travel times in Figure 1 are plotted assuming travelers have homogeneous preferences, allowing for heterogeneous $\beta/\alpha$ and $\gamma/\alpha$ while keeping their averages the same does not change the predictions for the peak travel time and length of rush hour.

Figure 1 shows three problems with matching the model to the data. The most significant problem is that the slope of the travel time profile is much lower than typically assumed.

In everyday speech “rush hour” refers to only the period of time when travel times are at their very worst; here I am using “rush hour” to denote the entire period of time when travel times are higher than in free flow conditions. It is this entire evolution of travel times across the day that structural congestion models are designed to explain.

With heterogeneous $\beta/\alpha$ and $\gamma/\alpha$ the model predicted travel times become convex on each side of the peak, as shown in Figure 2 of Arnott et al. (1994).
The maximum slope (over an hour interval) before the peak is 0.25, while the expected average slope is 0.6, and the maximum slope after the peak is −0.12, far below the expected average of −2.34. Of the five papers in Table 1 that allow for heterogeneous preferences, only one has a minimum inflexibility below the maximum seen in the data. In addition, the slope after the peak is more shallow than the slope before the peak: the ratio of the (maximum) slope after the peak to the (maximum) slope before the peak is 0.48 rather than near 4. Finally, the slope is relatively flat at the peak.

This mismatch between theory and data is not unique to California State Route 91. Figure 2 shows the pattern holds for I-55 in Chicago from I-80 to I-94 (67.8 kilometers, a long commute) and a more typical commute of I-290 from Wolf Road to I-94 (22.5 kilometers).\(^\text{10}\) So far all of this data is just for highways; however, it holds if we include surface streets as well. Figure 3 also plots travel times from Hopkinton to Boston Commons (48.5 kilometers, a long commute) and Wellesley to Boston Commons (24.3 km, a more typical commute) using data from Google Maps.\(^\text{11}\) In every case the slope before the peak is less than half that expected, the slope after is less than a seventh that expected, the ratio of the slope after the peak to the slope before the peak is always less than 2/5ths that expected, and the slope of the travel time profile is flat at the peak.

To further test whether there exist trips where travel times rise and fall in accordance with the theory, I use data on travel times for 4.4 million origin-destination pairs from 32 cities across the world from Uber Movement (2019). These travel times are the average weekday travel time for Uber trips for a given origin-destination pair for the third quarter of 2017, and are reported at the hourly level.\(^\text{12}\) For cities in the United States, origins and destinations are aggregated by Census tracts, and in other cities origins and destinations are aggregated in a roughly similar way. Appendix Table A1 lists the cities this data comes from as well as how many origins/destinations there are in each city.

Because the underlying data comes from Uber trips, this data will tend to include origin-destination pairs that are closer together or have either the origin or destination downtown or at an airport.

Using this data I calculate the largest and smallest observed slopes in the travel time profile between 4:00–11:59 AM. To avoid origin-destination pairs with little data, I require that there be observations for 6:00–10:59 AM. Figure 4 plots the distribution of observed largest $\beta/\alpha$ and largest $\gamma/\alpha$ implied by these slopes. The observed slopes are much lower than typically assumed. Out of 4.4 million origin-destination pairs, only 47 have a

\(^{10}\)Data is for all weekdays, and is from Illinois Department of Transportation (2017). The mean commute in the Chicago metropolitan area is 31.2 minutes (U.S. Census Bureau 2016).

\(^{11}\)Data covers all weekdays in July and August of 2017. The mean commute in the Boston metropolitan area is 30.2 minutes (U.S. Census Bureau 2016). Hopkinton, Wellesley, and Boston Commons are the start, mid, and endpoints of the Boston Marathon.

\(^{12}\)I use the third quarter as it minimizes the number of holidays, and 2017 as it maximizes the number of cities for which data is available.
maximum slope over 0.5, and only 3 have a slope over 0.6. Likewise for the minimum slope is much shallower than typically assumed. Furthermore, the smallest slope is typically of smaller or similar magnitude than the largest slope. The median ratio of $\gamma/\beta$ implied by the slopes is 0.80, and only 0.5 percent of the origin-destination pairs have above 3.9. Figure A1 plots the distribution of estimated $\gamma/\beta$.

4. Possible solutions

Having identified the three problems with structural models of congestion, I now explore possible solutions to them.

4.1. Slope of travel time profile too low. The most significant problem is that travel times rise and fall far slower than expected. This implies there are not any travelers who have even moderately inflexible schedules. However, Small (1982) reports that 63.4 percent of auto commuters in his sample cannot arrive late "without it mattering very much," and 58 percent of workers in the 2009 National Household Travel Survey report they cannot
choose what time they start work.\textsuperscript{13} These workers thus have a $\beta/\alpha$ near one (and a large $\gamma/\alpha$ as well). Inflexible travelers exist, they are just not affecting equilibrium travel times in the way the model with our standard assumptions predicts.

4.1.1. Preferred solution: Inframarginal travelers. My preferred solution to the problem of travel times climbing and falling too slowly is to allow for the possibility that some travelers are inframarginal and arriving exactly on-time. This works because, as (4) shows, when

\textsuperscript{13}While it is possible to reduce the percentage of travelers who are inflexible by only considering trips on the interstate in the morning and assuming all non-work trips are flexible, doing so still leaves 29 percent of travelers being very inflexible.
travelers are at a corner solution (i.e., arriving exactly on-time), their marginal rate of substitution need not be tangent to the budget line (i.e., travel time profile). As a result, we are only observing the inflexibility of those travelers who choose to arrive early or late. As the least inflexible travelers will choose to be early or late, the puzzle turns into the not particularly puzzling “the least inflexible travelers are not very inflexible.”

In order to have inframarginal travelers we need to have heterogeneous inflexibility and a continuum of desired arrival times. A continuum of desired arrival times is necessary because for a positive measure of travelers to arrive exactly on time, there must be a positive measure of desired arrival times. This modeling assumption, with otherwise homogeneous travelers, appeared in the initial papers using the bottleneck model (Vickrey 1969; Hendrickson and Kocur 1981), but was subsequently dropped as it did not affect equilibrium outcomes.\textsuperscript{14}

However, when travelers are heterogeneous, allowing for a continuum of desired arrival times does affect equilibrium outcomes. As Newell (1987) shows, adding this assumption means equilibrium travel times and tolls only depend on the preferences of some travelers, which allows us to rationalize our priors about traveler inflexibility with the empirically observed travel time profile.

\textsuperscript{14}To the best of my knowledge, the only exceptions are Newell (1987), de Palma and Lindsey (2002), and Hall (2018, 2019).
This is also analytically tractable. With otherwise homogeneous travelers it is possible to have fairly flexible specifications of desired arrival times and still find simple solutions (see Hendrickson and Kocur (1981)). Even with heterogeneous travelers it is possible to find closed-form solutions for the equilibrium both when the road is free and tolled, as while as for pricing a portion of the lanes, given the assumption that desired arrival times are uniformly distributed (see Hall (2018)).

Furthermore, the assumption of a continuum of desired arrival times is more reasonable than it initially sounds. While it is unlikely there is anyone who wants to arrive at work at 7:34:21.3, what matters for the model is when travelers want to arrive at the end of the bottleneck or highway, not when they want to arrive at work. Because the distribution of distances between the end of the bottleneck and work is continuous, the distribution of desired arrival times at the end of the bottleneck is also continuous.

This solution also rationalizes the difference between the empirical estimates from Small (1982) and the observed slope of the travel time profile. Small (1982) estimates the average inflexibility, but the slope of the travel time profile is driven by the inflexibility of those travelers who are not inframarginal.

4.1.2. Possible solution: Only measuring marginal inflexibility. Another possible solution is to note that we are just measuring marginal inflexibility at the equilibrium time of arrival. If
we generalize the schedule delay cost function $D$ to have more curvature, as in Lindsey (2004) and Fosgerau and Engelson (2011), then the first order conditions imply

$$T'(t) = \frac{D'_i(t^* - t)}{\alpha_i} \quad \text{if } t \neq t^*.$$  

In this case it is possible for travelers to really dislike being very early or very late, but to have low marginal inflexibility because they are actually arriving very close to their desired arrival time.

In order for this modeling change to resolve the puzzle, it must be combined with heterogeneity in inflexibility and desired arrival times. Without heterogeneous desired arrival times, it is still the case that median traveler arrives over an hour early or late, and, given the previous cited evidence on inflexible travelers, it seems implausible that the median marginal inflexibility for arriving so early or late is so low. Without heterogeneous inflexibility it is difficult to explain the choices of those travelers who arrive exceptionally early or late. With heterogeneous inflexibility, the inflexible arrive near the peak and on the margin are quite flexible, while the flexible arrive further from the peak and are, likewise, on the margin quite flexible.

This is conceptually similar to allowing for inframarginal travelers. In both cases inflexible travelers exist, but their inflexibility is not reflected in equilibrium travel time profiles either because they are arriving exactly on-time, or close enough to on-time that on the margin they are flexible.

While generalizing the schedule delay cost function does solve the problem, I prefer to allow for inframarginal travelers for three reasons. First, allowing for inframarginal travelers is more parsimonious. Both solutions require heterogeneity in inflexibility and desired arrival times. While generalizing the schedule delay cost function allows the heterogeneity in desired arrival times to be discrete rather than continuous, it also requires the further step of generalizing the schedule delay cost function. Second, and as a direct consequence, allowing for inframarginal travelers is more tractable. Third, it matches our intuition, and survey evidence, that many people strictly prefer their chosen arrival time over all others. Small (1982) finds that 35.5 percent of auto commuters in the San Francisco Bay Area arrive on-time, and thus are likely inframarginal. Hall (2019) reports that 43 percent of travelers using California State Route 91 say they do not typically leave early or late to avoid traffic; this suggests they are (typically) inframarginal.

4.1.3. Possible solution: Change parameter values. While the solutions in Sections 4.1.1 and 4.1.2 solve the problem, they require adding heterogeneous preferences. There are many questions for which models with homogeneous travelers are appropriate, in these situations using small values for $\beta/\alpha$ and $\gamma/\alpha$ (such as 0.22 and 0.20, the averages of those reported in Figures 1–3) will allow the model to give reasonable predictions for travel times.

15Heterogeneous inflexibility in this case means that $D'_i(x)/\alpha_i \neq D'_j(x)/\alpha_j$ for $i \neq j$. 


4.1.4. **Possible solution: All travelers are inframarginal.** An extreme special case of allowing for inframarginal travelers is to assume that all, or nearly all, travelers are inframarginal with respect to the choice of when to arrive, meaning all, or nearly all, travelers are arriving at their desired arrival time. As discussed in Section 4.1.1, this requires allowing for a continuum of desired arrival times and means the travel time profile only gives bounds on traveler preferences. This solves all three problems, as in equilibrium the travel time profile could be relatively flat, leading everyone (or almost everyone) to conclude it is not worthwhile to leave early or late.\(^{16}\)

That said, this is implausible for medium and large cities. Hall (2019) reports that 57 percent of those using California SR-91 actually leave early or late to avoid traffic, which means a majority of travelers are marginal with respect to the choice of when to arrive. It also seems unlikely that many travelers have a desired arrival time of 6 AM, and so the travelers doing so are likely marginal.

For small cities, this may very well hold. In these cities travel times are higher at the peak of rush hour, but perhaps not by enough to induce anyone to leave early or late.

4.1.5. **Possible solution: Long- vs. short-run preferences.** Another possible solution is to note that long- and short-run preferences can differ, as in Peer et al. (2015). In the long run travelers can adjust on more margins of behavior, and so are more flexible. Consider, for example, how easy it can be to arrange your schedule to arrive an hour later than usual versus the cost of arriving an hour late unexpectedly. If our intuition for reasonable parameter values is based on thinking about the short run, then we will be looking for values of inflexibility that are too large.

While this helps, as discussed above, many workers lack the ability to choose their start time, even in the long run. Workers with fixed start times remain very inflexible even after making the distinction between long- and short-run preferences, and this solution does not explain why they are not affecting equilibrium travel times as predicted.

4.1.6. **Possible solution: Add value of reliability.** A sixth possible solution is to consider the value of reliability by adding in uncertainty over travel times. As Figure 6 shows, the amount of uncertainty in travel times varies over rush hour. Adding this into structural congestion models, as in Arnott et al. (1999) and Fosgerau and Karlström (2010), would provide an extra incentive for travelers to leave early or late—travelers could reduce the uncertainty in their travel time.

\(^{16}\)In the bottleneck model this requires that the density of desired arrivals at any point during rush hour exactly equals road capacity, and causes travel times to be indeterminate (that is, there is not a unique equilibrium). The model of Henderson (1974) and Chu (1995) has a unique equilibrium when using this solution. In this model, congestion only depends on the number of travelers arriving at a given time. Travel times rise and fall based on the demand for a given arrival time, without necessarily requiring any travelers to arrive early or late.
There are two reasons this fails to solve the problem. First, a model with uncertainty predicts that travel times at the start and end of rush hour should climb at the same rate they do in a model without uncertainty, and thus even with adding uncertainty some other solution must explain the low slopes of the empirical travel time profile at the start and end of rush hour.

To see this, consider a slightly transformed model that includes uncertainty. In contrast to above, I now focus on departure times rather than arrival times. Assume there is a distribution of possible states of the world, $F(\phi)$, where a larger $\phi$ means worse traffic, $\partial T(t_d, \phi)/\partial \phi > 0$; and that leaving later means arriving weakly later, $\partial T(t_d, \phi)/\partial t_d \geq -1$. Define $\hat{\phi}(t_d)$ as the largest $\phi$ so that a traveler departing at $t_d$ arrives on-time.

Re-writing the cost function in terms of departure times gives

$$p(t_d, \phi) = \alpha T(t_d, \phi) + D(t^* - t_d - T(t_d, \phi)).$$

(5)
Travelers choose their departure time to minimize their expected trip cost, which yields the following first order condition:

\[
\frac{dE(p(t_d, \phi))}{dt_d} = \alpha \int_0^\infty \frac{\partial T(t_d, \phi)}{\partial t_d} dF(\phi) \]

\[
- \beta \int_0^{\phi(t_d)} \left(1 + \frac{\partial T(t_d, \phi)}{\partial t_d}\right) dF(\phi) \]

\[
+ \gamma \int_{\phi(t_d)}^\infty \left(1 + \frac{\partial T(t_d, \phi)}{\partial t_d}\right) dF(\phi) = 0. \tag{6}
\]

For travelers who are always early, so that \(F(\hat{\phi}(t_d)) = 1\), (6) simplifies to

\[
\frac{dE(T(t_d, \phi))}{dt_d} = \frac{\beta_i}{\alpha_i - \beta_i}, \tag{7}
\]

while if a traveler is always late, so that \(t_d > t^*\) and \(F(\hat{\phi}(t_d)) = 0\), (6) simplifies to

\[
\frac{dE(T(t_d, \phi))}{dt_d} = -\frac{\gamma_i}{\alpha_i + \gamma_i}. \tag{8}
\]

The slopes for always early and always late travelers are identical to what we obtain in a world without uncertainty. The right-hand sides of equations (7) and (8) are not the same as (2) and (3) because (2) and (3) are in terms of arrival time rather than departure time. For the derivation of the slope of the travel time profile in terms of departure times, which matches (7) and (8), see Arnott et al. (1993).

Therefore, even in a model with uncertainty, the slope at the start and end of rush hour should be much steeper than we observe in the data, and so adding uncertainty over travel times does not resolve the problem.

Second, the magnitudes are such that while adding uncertainty helps narrow the gap between the parameters that seem reasonable and model predictions during the middle portions of rush hour, it does not close it. Fosgerau and Karlström (2010) estimate that the value of reliability accounts for about 15 percent of total time costs while Small et al. (2005) estimates it accounts for about a third of total time costs. The slope of the travel time profile is climbing roughly half as fast as theory suggests, so to close this gap the value of reliability would need to account for 50 percent of total time costs.

It is worth noting there are many possible sources of travel time uncertainty. These include demand fluctuations, as well as capacity shocks due to construction, weather, accidents (as in Fosgerau and Lindsey (2013)), or endogenous capacity breakdown (as in Hall and Savage (2018)). The analysis above covers all of these cases.

4.1.7. Possible solution: Heterogeneous trip lengths and origin-destination pairs. A final possible solution is to consider different trip lengths and origin-destination (O-D) pairs. Any given road segment is used by travelers with a variety of O-D pairs, and so equilibrium travel times for any given O-D pair are the result of the choices of travelers with a variety of
different O-D pairs. As such, it is worth considering whether a model that accounted for
the variety of O-D pairs would change the relationship between traveler preferences and
the slope of the travel time profile.

However, considering heterogeneous O-D pairs does not solve the problem. While it is
possible, and even likely, that for a given O-D pair there is not someone arriving at every
point in time, it is still the case that if a traveler is arriving early or late, then the slope of the
travel time profile at the point in time he arrives must be tangent to his indifference curve.
In the data we never see slopes like those we would expect given reasonable parameters,
and thus considering different trip lengths and O-D pairs does not solve the problem.

4.2. **Slope after peak too small relative to before peak.** The second problem is that the
slope of the travel time profile after the peak is too small relative to the slope before the
peak. By (2) and (3), the ratio of the slope before and after is $\gamma / \beta$, and measures the cost of
being late relative to being early. This ratio being small in absolute value suggests travelers
do not mind arriving late relative to the cost of being early.

4.2.1. **Preferred solution: Many travelers do not mind arriving late.** One interpretation of this
result is that the marginal traveler who is late incurs lower schedule delay costs than
the marginal traveler who is early. Furthermore, recalling the distinction between long-
and short-run preferences, being late does not necessarily mean literally arriving late to
an appointment, but can mean a traveler prefers to go to the doctor at 9 AM but instead
schedules the appointment for 11 AM to avoid traffic. He arrives exactly on-time to his 11
AM appointment, but still has schedule delay costs. In addition, everyday experience with
the frequency of late arrivals implies many people prefer arriving late to arriving early, at
least for many non-work trips.

There are three ways to add travelers who do not mind being late to the model. The first
is to allow for heterogeneity in $\gamma / \beta$ so that some travelers find being late very costly, while
others will not mind being late. Those who find it costly will arrive early, while those who
do not mind being late will arrive late. For the bottleneck model, Arnott et al. (1994) shows
how to solve for equilibrium analytically when this is the only source of heterogeneity in
traveler preferences and Liu et al. (2015) shows how to solve for equilibrium numerically
when there are additional dimensions of heterogeneity.

A second is to explicitly model the difference between the long and short-run preferences,
as in Peer et al. (2015).

Finally, the simplest approach is to simply use a low value for $\gamma / \beta$. While this amounts
to assuming there are no travelers who hate arriving late, this only matters if in the
counterfactual policy there are travelers who switch from arriving early to arriving late, or
vice-versa. If the policy does not change the fraction of travelers who are late (conditional
on other preference parameter values), then this assumption simplifies analysis without
compromising results. As this is tractable without compromising results, this is my preferred solution.

4.2.2. **Possible solution: Add value of reliability.** Another possible solution is adding the value of reliability by allowing for uncertainty in travel times. If there are no travelers who are always late, then the slope after the peak can be less steep than in the deterministic model.

The intuition for this is that when a traveler who is sometimes early and sometimes late, makes a small delay in their departure time, they decrease their expected schedule delay early (a good thing) while increasing their expected schedule delay late (a bad thing). Thus for a given change in departure time, expected schedule delay costs change by less than they would in a model without uncertainty. In order to keep travelers on their first order condition, this means expected travel times likewise change by less than they would in a model without uncertainty.

This result for the bottleneck model is implicit in Proposition 1 of Arnott et al. (1999). To derive it with more general assumptions on the congestion technology we add the assumption that $T(t_1, \phi_1) > T(t_2, \phi_1) \iff T(t_1, \phi_2) > T(t_2, \phi_2)$. This implies that if we exogenously increase the travel time at $t_d$ for one $\phi$, we have increased it for all $\phi$, and so for any set $\Phi \subset \mathbb{R}$

$$\frac{dE\left(\frac{\partial T(t_d, \phi)}{\partial t_d} \bigg| \phi \in \Phi\right)}{dE\left(\frac{\partial T(t_d, \phi)}{\partial t_d}\right)} > 0.$$  

Next, note $\hat{\phi}(t_d)$ is implicitly defined by

$$t_d + T(t_d, \hat{\phi}(t_d)) = t^*,$$

and by the implicit function theorem

$$\frac{d\hat{\phi}(t_d)}{dt_d} = -\left(1 + \frac{\partial T(t, \hat{\phi}(t_d))}{\partial t_d}\right) \left(\frac{\partial T(t, \hat{\phi}(t_d))}{\partial \phi}\right)^{-1} \leq 0.$$  

By the implicit function theorem and (6)

$$\frac{d^2E\left(T(t_d, \phi)\right)}{dt_d^2} = (\beta + \gamma) \left[1 + \frac{\partial T}{\partial t_d}\right] \frac{d\hat{\phi}}{dt_d} \times \left[(\alpha - \beta)F\left(\hat{\phi}(t_d)\right) \frac{dE\left(\frac{\partial T(t_d, \phi)}{\partial t_d} \bigg| \phi \leq \hat{\phi}(t_d)\right)}{dE\left(\frac{\partial T(t_d, \phi)}{\partial t_d}\right)} + (\alpha + \gamma) \left[1 - F\left(\hat{\phi}(t_d)\right)\right] \frac{dE\left(\frac{\partial T(t_d, \phi)}{\partial t_d} \bigg| \phi > \hat{\phi}(t_d)\right)}{dE\left(\frac{\partial T(t_d, \phi)}{\partial t_d}\right)}\right]^{-1} \leq 0. \quad (9)$$
Thus expected travel times are concave, and there are more than two slopes. Combining this with (8) implies that if the last traveler to arrive is not always late, then the slope of the travel time profile after the peak is less than it is in the model without uncertainty.

However, there is a possible contradiction that may prevent adding uncertainty from solving the problem of the slope after the peak being too small relative to the slope before the peak.\textsuperscript{17} It seems unlikely that there are no travelers who are always late unless the cost of being late is very large. But if the cost of being late was large, the slope after the peak of rush hour would be steep. Thus it seems unlikely that adding uncertainty resolves the problem of the slope after the peak being small relative to the slope before the peak.

4.2.3. **Possible solution: All travelers are inframarginal.** As discussed in Section 4.1.4, if all travelers are inframarginal then the first-order conditions imply that the travel time profile only provides bounds on travelers’ preferences. However, as discussed earlier, the data suggests many travelers are marginal.

4.3. **Slope flat at peak.** The third problem is that the slope is relatively flat at the peak of rush hour, while the model implies it should reach a sharp peak. Fortunately, after fixing the first and second problems, the magnitude of error introduced by this third problem error is small. To see this, consider drawing the triangle that best fits Figures 1, 2, or 3. While it misses the flatness of the peak, it matches significantly better than the model does under the current assumptions.

All of the solutions explored below work, and so my recommended solution will solely be on the grounds of tractability.

4.3.1. **Preferred solution: Only measuring marginal inflexibility.** Generalizing the schedule delay cost function, as in Section 4.1.2, also solves this problem. This allows for travelers to not find it particularly costly to arrive 10 minutes early or late while hating being an hour early or late. For a more generic schedule delay cost function with homogeneous travelers to explain the travel times in Figures 1, 2, and 3 requires assuming travelers do not mind arriving thirty minutes early or late. This is implausible. However, with heterogeneous desired arrival times it is possible for the window of indifference to be much smaller.

A similar solution would be to keep the standard piecewise linear schedule delay costs, but add the assumption that travelers are indifferent between arrivals in a small window around their desired arrival time. As with generalizing the schedule delay cost function, in order to have a reasonably small window of indifference we must combine it with heterogeneous desired arrival times. As this is the most tractable of all the solutions to the third problem, it is my preferred solution.

\textsuperscript{17}It is a “possible” contradiction because for adding uncertainty to help, it must be the case that no travelers are always late, and for this to be plausible, travelers must have heterogeneous desired arrival times. However, a structural congestion model has yet to be solved with uncertainty in travel times and heterogeneous desired arrival times.
4.3.2. **Possible solution: Add value of reliability.** Adding the value of reliability by allowing for uncertainty solves this problem. As implied by (9), near the peak, small changes in departure time can lead to imperceptible changes in expected schedule delay costs as decreases in expected schedule delay early are offset by increases in schedule delay late, or vice versa. This leads to travelers being relatively indifferent over a variety of arrival times near the peak, and thus a relatively flat peak.

4.3.3. **Possible solution: All travelers at peak are inframarginal.** Again, allowing for inframarginal travelers can help. A less strict version of the solution in Section 4.1.4, would be to assume all, or almost all, those arriving at the peak of rush hour are inframarginal with respect to the choice of when to arrive. As before, this requires a continuum of desired arrival times so a positive measure of travelers can arrive exactly on-time.

In contrast to the other possible solutions, this solution does depend on the congestion technology. For (almost) all those arriving at the peak to be inframarginal, while those arriving further from the peak are not, it must be that (almost) everyone who desires to arrive at the peak is able to do so, while not everyone who desires to arrive further from the peak is able to do so. This either requires demand to be higher further from the peak, which seems implausible, or supply to be higher at the peak. That is, the congestion technology must allow for greater throughput when travel times are high. This could happen either because of the traditional trade-off between throughput and speed on a single route, or because, as Akbar and Duranton (2017) argue, as travel times climb, additional routes are used, and so system capacity increases.\(^\text{18}\)

Combining a continuum of desired arrival times with a congestion technology that allows greater throughput when travel times are high solves the problem of the flat peak because it allows almost everyone who wishes to arrive at the peak of rush hour (say, 8–9 AM) to do so. Thus, at the peak of rush hour the marginal traveler who is early or late is very flexible, and so the slope of the travel time profile is very flat. In contrast, at other times system capacity is lower and so more travelers must arrive early or late. The marginal traveler who is arriving is less flexible and so the slope is steeper.

5. **Why these problems matter**

These failures of the standard structural congestion models to fit the data limit our ability to use these models to quantify magnitudes and change qualitative results with heterogeneous travelers. In this section I give examples of both.

Both of the examples use the congestion technology of the bottleneck model. In the bottleneck model, travel along the road is uncongested, except for a single bottleneck through which at most \(s\) vehicles can pass per unit time. When the departure rate, \(\rho(t),\)

\(^{18}\) This point is implicit in models of route choice when the routes do not have identical free flow travel times, such as the models of Pigou (1920), Arnott et al. (1990), and Liu and Nie (2011).
exceeds $s$, a queue develops. Denoting queue length as $Q(t)$, variable travel times are given by, $T^v(t) = Q(t)/s$.

As an example of how our model’s failure to match the data using reasonable parameter values affects our ability to quantify magnitudes, consider the task of measuring the social welfare gains from congestion pricing. In the bottleneck model with homogeneous travelers the social welfare gains from adding first best tolls equal the total variable travel time multiplied by the value of time (Arnott et al. 1993). As Figure 1 vividly demonstrates, our model fails to accurately predict total variable travel time, and as a result, we overestimate the welfare gains by a factor of 23.19

Furthermore, inasmuch as these empirical problems imply the standard structural models of congestion are wrong in important ways, and inasmuch as changing our model changes our qualitative results, then these empirical problems change our qualitative results. As an example, consider the distributional effects of congestion pricing with heterogeneous travelers. Adding inframarginal travelers by assuming a continuum of desired arrival times adds travelers who strongly prefer their current arrival times. A policy that changes when they arrive hurts these travelers significantly. Moreover, as Hall (2018) shows, adding inframarginal travelers shrinks the set of parameter values for which congestion pricing generates a Pareto improvement.

6. Conclusion

Structural models of traffic congestion are a powerful tool for answering a wide variety of important questions. Improving these models will help us answer these questions better.

This paper makes three contributions. First, it shows that while these models generate a travel time profile that has the right general shape, they fail to match observed travel times. I identify three problems with matching these models to the data: (1) travel times climb and fall far slower than the models predict, (2) travel times fall too slowly after the peak relative to how quickly they climb before the peak, and (3) travel times are essentially flat at the peak. These problems are a consequence of how we model traveler preferences, rather than the model of congestion itself, and so affect the bottleneck model, cell transmission model, hydrodynamic traffic flow model, and the models of Henderson (1974) and Chu (1995).

Second, it shows that while we do not expect models to be perfect representations of reality, these problems matter because they affect quantitative results with homogeneous or heterogeneous travelers, as well as qualitative results with heterogeneous travelers. Thus

19 In Figure 1, total variable travel time is the area under the curve. Arnott et al. (1993) show the welfare gains equal $(1/2)(N^2/s)[\beta\gamma/(\beta + \gamma)]$. I use this equation to calculate the welfare gains using (1) the standard assumptions for $\beta/\alpha$ and $\gamma/\beta$, and choosing $N/s$ to match (a) the length of rush hour or (b) peak travel time, and (2) choosing $\beta/\alpha$, $\gamma/\beta$, and $N/s$ to match the length of rush hour, peak travel time, and when the peak occurs. In the text I report the comparison of (1a) to (2), if instead we compare (1b) to (2) we underestimate the welfare gains by a factor of 0.55.
there is a need for more realistic assumptions that improve our ability to meaningfully answer a wide variety of questions.

Third, it explores a variety of possible solutions, and makes the following recommendations for improving the fit of structural models of traffic congestion. If the goal is to quantify some outcome with homogeneous travelers, then it is easier to simply use significantly lower parameter values for travelers’ schedule delay costs than are typically assumed. If the goal is a quantitative or qualitative result with heterogeneous travelers, then allow for inframarginal travelers by assuming travelers’ desired arrival times are continuously distributed and travelers differ in the flexibility of their schedules. Furthermore, use parameter values that imply the cost of being late is less than the cost of being early. These solutions do not fix the problem of travel times being essentially flat at the peak, but the magnitude of this problem is small after fixing the other two. If it must to be fixed, this is most tractably done by allowing travelers to be indifferent between arrivals in a small window around their desired arrival time.

References


Table A1. Uber Movement sample composition

Notes: Data from Uber Movement 2019. This table reports the composition of the sample used in Figures 4 and A1.

Appendix A. Additional tables and figures

Figure A1. Estimated distribution of $\gamma / \beta$ for 4.4 million origin-destination pairs. Data from Uber Movement 2019.

Notes: To save space, this figure does not plot the 0.1 percent largest and smallest observations.