Bottleneck Congestion: The Welfare Cost of Morning Commutes

Zach J. L. Lee*

Nanyang Technological University, Singapore

2025-08-01

Abstract. Commuting by car remains central to daily life for many, and the morning rush to meet similar work arrival times creates queuing externalities as drivers compete for limited road capacity. I use the United States National Household Travel Survey (2001, 2009, 2017) to estimate average morning-peak queueing delays under the bottleneck model. Controlling for trip distance, neighbourhood, and individual covariates, I find delays rose from 2.84 minutes in 2001 to 3.25 minutes in 2017. The 2017 average delay yields a lower bound annual welfare loss of $11.8 billion. Cross-MSA heterogeneity shows queueing delays increase with income and vehicle miles travelled but decline with toll-lane expansions. Building on the bottleneck framework, I propose a second-best "coarse" block toll, calibrated to the observed delay profile, to partially reclaim commuter welfare. I discuss its implementation alongside New York City's 2025 cordon-based congestion pricing. These findings underscore both the persistence of congestion costs and the potential gains from targeted pricing strategies.

*I thank Jinwon Kim for sharing original materials that shaped the early stages of this paper. I am grateful to Euston Quah, Jiaoying Pei, and Jonathan H. W. Tan for insightful comments. I also thank Daniel Jenkins, Robert Rozycki, and Roff Thomas from the Federal Highway Administration for guidance on the publicly available datasets.

1 Introduction

Commuting by private vehicle remains a daily ritual for most workers in the United States. Even in 2022, during the COVID-19 pandemic, about 41.3 billion commute trips by vehicles were made, dropping only 28% compared to 2017 (Bricka et al., 2024). Private-vehicle commutes accounted for 90.5% of all commutes. The morning rush to meet common arrival-at-work times produces distinct, persistent peak-period congestion. Schrank et al. (2024) report that while overall morning congestion fell during the pandemic, the morning peak itself remained, contrasting with whole-day congestion patterns that shifted markedly.

This paper provides a nationally representative, trip-level estimate of morning-peak queueing delays for 2001, 2009 and 2017 using the canonical bottleneck framework. A reduced-form approach recovers route-integrated congestion queueing time without specifying a travel-demand curve, and maps directly to policy metrics such as morning-peak welfare loss and time-varying congestion-pricing profiles. I estimate an average queueing delay of 3.25 minutes in 2017. Aggregated, this time loss, imposed by and across commuters, mounts to a conservative welfare loss of $11.8 billion. Beyond providing a policy-relevant measure of social cost, I discuss how the estimated average queueing-time profiles can guide practical time-block pricing, in relation to New York's congestion-pricing scheme introduced in 2025.

Pigou (1920) first highlighted that congestion imposes a Pigouvian externality on road users. Measuring these welfare losses is challenging as one must simultaneously recover travel demand and cost functions. Empirical applications of the Pigouvian framework typically recover the inverse speed–density (i.e., travel-time cost) function on specific corridors and then infer marginal external congestion costs (supply curves) using traffic-count (Dewees, 1979; Fosgerau & Small, 2012; Walters, 1961). Yang et al. (2020) extend this approach to estimate a city-wide cost function for Beijing, instrumenting density with driving-restriction policies. By not explicitly estimating the travel demand function, these corridor-based methods cannot recover the true Pigouvian deadweight loss and rely instead on assumed demand curves. They also use segment-level speed counts that do not fully capture a trip's cumulative queueing delays along its entire route.

Vickrey (1969) describes an alternative framework that encapsulates such a queueing delay. Called the bottleneck model, this was further developed and formalised in a series of

papers by Arnott, de Palma & Lindsey, hereafter referred to as ADL (Arnott et al., 1990, 1993, 1994).[1] In the bottleneck model, commuter timing decision is endogenous. Travellers trade off in-vehicle time and early–late arrival costs when capacity has been exceeded. Under piecewise-linear congestion formation and commuter cost, together with homogeneous commuters sharing a common desired arrival time, the model yields a closed-form expression for the commuter's queueing time.[2] The key primitives in this expression are the value of in-vehicle time, schedule-delay costs associated with not arriving on time, and the demand–capacity imbalance. While studies have measured time-related costs (Kim & Moon, 2024; Small, 1982), estimates are typically geographically narrow or temporally limited. The demand–capacity imbalance is harder still to measure because it requires commuter counts and effective capacity across many routes.

To estimate bottleneck queueing time, Kim (2019) applies a reduced-form strategy to National Household Travel Survey (NHTS) data to recover nationwide average queueing delay. Multiplying by an appropriate value of time then yields the social cost of congestion. Kim also recovers average queueing time for large Metropolitan Statistical Areas (MSAs)[3] and studies determinants of congestion by pairing NHTS trips with road-stock data from the Highway Performance Monitoring System (HPMS).

Related work uses trip-level and real-time data. Couture et al. (2018) model urban travel as a production technology and estimate city-level supply functions, while still relying on assumed demand to compute deadweight loss. Akbar et al. (2023) combine counterfactual travel times from Google Maps with observed trip surveys to identify demand, showing across 180 Indian cities that cross-city speed differences largely reflect free-flow conditions rather than peak-period queueing. Cook et al. (2025) pair Google Maps data with a synthetic-control design to evaluate congestion pricing in Manhattan. Kim and Moon (2024) combine NHTS trips with high-frequency Google Maps travel-time profiles in California to recover heterogeneous scheduling parameters and route-specific queueing delays. Kreindler (2024) deploys a GPS-based field experiment in Bangalore to elicit scheduling preferences under endogenous peak-hour queues and quantify the welfare effects of dynamic

[1] Small (2015) gives a concise review of the bottleneck model. An exhaustive list of research related to the bottleneck model is available in Li et al. (2020).

[2] Small and Verhoef (2007, pp. 88–93) provides a general treatment of the bottleneck model without these assumptions.

[3] A MSA is defined by the White House Office of Management and Budget as a core urban area of at least 50 thousand population and adjacent territories with strong social and economic ties. In the NHTS, large MSAs have a population of more than 1 million.

tolls within the bottleneck framework. These studies push measurement to the route and minute, but most either remain city- or corridor-specific, or rely on assumed demand to monetise losses.

By contrast, I use nationally representative trip diaries to recover route-integrated morning-peak queueing delay without specifying demand, and I extend Kim's single-year approach to three repeated cross-sections (2001, 2009, 2017) to track evolution over time. This delivers two outputs that are directly comparable across place and time: Average Queueing Cost (AQC) and Total Queueing Cost (TQC). The estimates complement corridor-based Pigouvian methods by providing commute-level deadweight in the bottleneck model and complement real-time Google-Maps/GPS evaluations by delivering a national and metropolitan baseline against which policy pilots like New York's 2025 congestion pricing programme can be interpreted.

Building on Kim's single-year (2009) analysis, I extend the method to three repeated cross-sections (2001, 2009, 2017) to trace the evolution of morning-peak delays alongside changes in demographics, commuting patterns, transport infrastructure and the built environment. For 2017, I compute Average Queueing Cost (AQC) and Total Queueing Cost (TQC) at national and MSA scales. The study complements corridor-based Pigouvian approaches by estimating route-integrated queueing delays from trip data, and it provides calibrated toll profiles consistent with the bottleneck model. In sum, I offer a tractable and representative estimate of peak-period commuting inefficiency costs. Together, these features provide a consistent, scalable yardstick for valuing peak-period welfare losses and comparing congestion-pricing designs across cities and years.

Theoretical and empirical frameworks for this paper are presented in Section 2 and Section 3. Section 4 describes the NHTS, HPMS data and sample construction. Section 5 reports national- and MSA-level estimates of average queueing delays, and analyses their structural determinants. Section 6 quantifies individual and aggregate welfare losses, and discusses congestion pricing alongside complementary strategies to mitigate them. Section 8 concludes.

2 The bottleneck model

2.1 Bottleneck congestion technology

I review the canonical fixed commuters bottleneck model in its most general form, as described by ADL (Arnott et al., 1990, 1993), using Kim's notation throughout.

There are N identical morning commuters who need to reach their work destination via a route with a single bottleneck. The bottleneck has a fixed capacity of ψ cars per unit time with a first-in-first-out queueing system.

Let $r(t)$ denote the arrival rate, in cars per unit time, into the bottleneck queue at time t, and $q(t)$ queue length. At each time instant u, the queue grows or shrinks by exactly the excess of arrivals over departures from the bottleneck. If $r(u)$ cars arrive and the bottleneck processes ψ cars at the same instance, the net addition to the queue is $\Delta q = r(u) - \psi$. Summing these instances of net arrivals from the moment the queue appears up to time t results in $q(t)$. Thus, the rate of change of queue length is given by:

$$\dot{q}(t) = r(t) - \psi \quad \text{when } q(t) > 0. \tag{1}$$

A car arriving when $q(t)$ vehicles are already waiting must wait a queueing delay of

$$Q(t) = q(t)/\psi. \tag{2}$$

If the commute route has a distance x, and speed under free-flow conditions is normalised to one, the free-flow or non-congested travel time along the route is simply x.[4] Putting these together, I get a total travel time of

$$Y(t) = x \ + \ Q(t). \tag{3}$$

[4]In expositions of the bottleneck model, including ADL, it is common to set the free-flow time to zero for convenience without the loss of generality. Here, x is explicitly needed for the estimation approach used in this paper.

I assume that each of the N commuters are driving their own cars. If I denote $[t_0, t_1]$ as the arrival times at the bottleneck queue for the first and last of these N commuters, flow conservation necessitates that the time period with a queue at the bottleneck is given by

$$t_1 - t_0 = \frac{N}{\psi}. \tag{4}$$

This canonical bottleneck technology provides the building blocks for a dynamic equilibrium model in which the duration and intensity of the morning peak are endogenous, and individual travel delays changes continuously over time (Small & Verhoef, 2007, p. 88). In the next section, I embed this queueing framework into each commuter's cost-minimization problem, trading off queueing delay against early- and late-arrival penalties.

2.2 Commuter cost

Each commuter chooses arrival time t at the bottleneck queue to minimise her private cost, which trades off:

1. Queueing cost $\quad a\,Q(t)$,
2. Early-arrival penalty $\quad b\,[\,t^* - t - Y(t)\,]_+$,
3. Late-arrival penalty $\quad c\,[\,t + Y(t) - t^*\,]_+$,

where t^* is the desired arrival time at the work destination, a is the shadow price of commute time, b and c are the marginal dis-utilities of early and late arrival respectively, and $[\cdot]_+$ denotes $max[0, \cdot]$. Early and late arrivals are collectively known as schedule delays. As with ADL, I assume that c > a > b in accordance with empirical results from Small (1982). Thus, the linear cost private trip cost of a commuter is given by:

$$C(t) = a\,Q(t) + b\,[\,t^* - t - Q(t)\,]_+ + c\,[\,t + Q(t) - t^*\,]_+. \tag{5}$$

ADL define equilibrium as a pure-strategy Nash outcome in which arrival time at the bottleneck is the strategy. Holding others' arrivals fixed, no commuter can lower her $C(t)$ by changing her own arrival (i.e., $C(t)$ must be constant on the queue-forming interval $[t_0, t_1]$). Hence

$$a\,\dot{Q}(t) \;+\; b\,\frac{d}{dt}[\,t^* - t - Q(t)\,] \;+\; c\,\frac{d}{dt}[\,t + Q(t) - t^*\,] = 0, \tag{6}$$

$$\tfrac{d}{dt}[\,t^* - t - Q(t)\,] = -1 - \dot{Q}(t) \quad \text{when } t < t^* \tag{7}$$

$$\tfrac{d}{dt}[\,t + Q(t) - t^*\,] = 1 + \dot{Q}(t) \quad \text{when } t > t^*. \tag{8}$$

From Equation 6, Equation 7, and Equation 8, I can obtain the equal-trip-price travel delay slopes:

$$\dot{Q}(t) = \begin{cases} \dfrac{b}{a-b}, & t \in [t_0, t^*], \\ -\dfrac{c}{a+c}, & t \in [t^*, t_1]. \end{cases} \tag{9}$$

From Equation 1 and Equation 2, queueing time evolves by:

$$\dot{Q}(t) = \frac{r(t) - \psi}{\psi}. \tag{10}$$

Combining Equation 9 and Equation 10 gives me two constant-rate regimes for arrival at the bottleneck queue that satisfies the equal-trip-price condition:

$$r(t) = \psi \;+\; \psi\,\dot{Q}(t) = \begin{cases} \dfrac{a\,\psi}{a-b}, & t \in [t_0, t^*], \\ \dfrac{a\,\psi}{a+c}, & t \in [t^*, t_1]. \end{cases} \tag{11}$$

The first and last commuters incur only schedule delays, so $Q(t_0) = Q(t_1) = 0$. Equivalently,

$$b\,(t^* - t_0) = c\,(t_1 - t^*). \tag{12}$$

Solving Equation 4 and Equation 12 jointly yields:

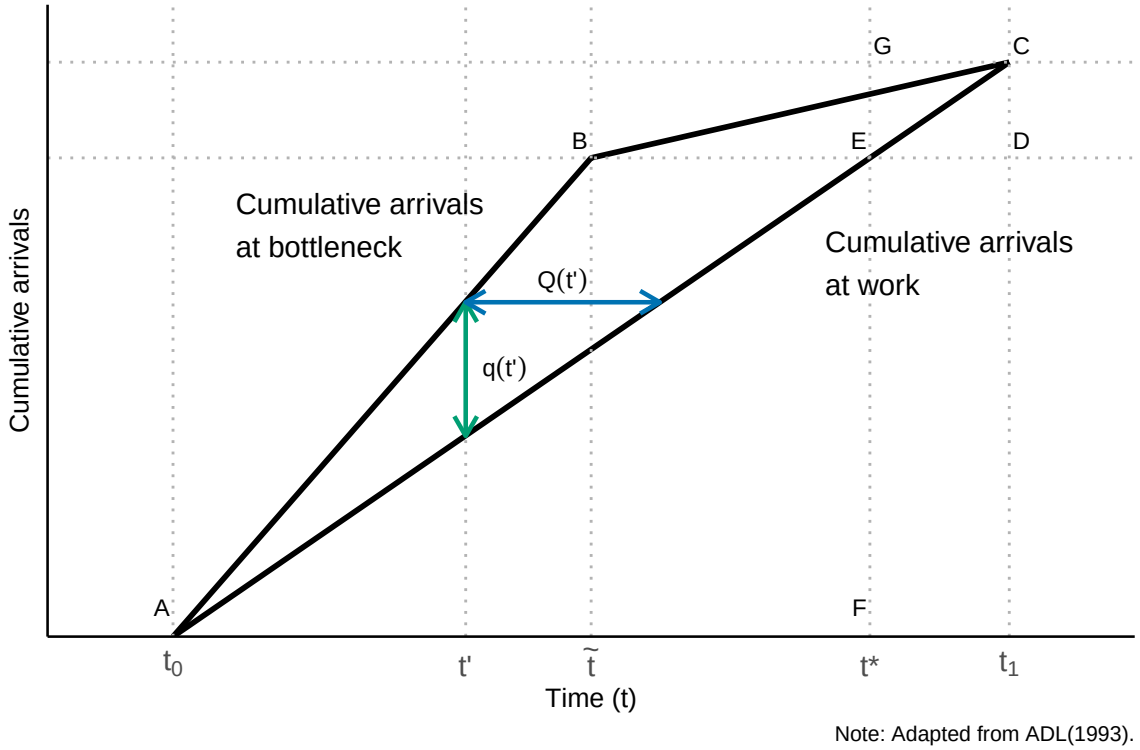


Figure 1: Cumulative arrivals at bottleneck queue and at work over time.

$$t^* - t_0 = \frac{c}{b+c}\frac{N}{\psi}, \tag{13}$$

$$t_1 - t^* = \frac{b}{b+c}\frac{N}{\psi}. \tag{14}$$

Similarly, it can be shown that

$$t^* - \tilde{t} = \frac{b}{a}\frac{c}{b+c}\frac{N}{\psi} \tag{15}$$

where $\tilde{t}$ is the arrival time at the bottleneck that incurs maximum queue length but no schedule delay.

In Figure 1, the vertical distance (in green) between the cumulative arrival schedules at the bottleneck and at work is the queue length, $q(t')$. The horizontal distance (blue) between the schedules is the queueing time, Q(t'). Both build up linearly to $\tilde{t}$ and fall off linearly thereafter. Maximum queueing time is incurred at the desired arrival at work time, t^*.

The total queueing cost, TQC, is price a multiplied by triangle ABC, the total queueing time. The total schedule delay cost, SDC is the sum of b multiplied by triangle AEF (total time units of early arrivals away from t^*), and c multiplied by triangle CEG (total time units of late arrivals away from t^*). The total cost, TC, is $TQC + SDC$. The average bottleneck total cost, ATC, is the cost incurred by each commuter (Equation 5) due to the bottleneck, and is just TC divided by N. Solving for TQC and SDC using Equation 13, Equation 14, and Equation 15 gives:

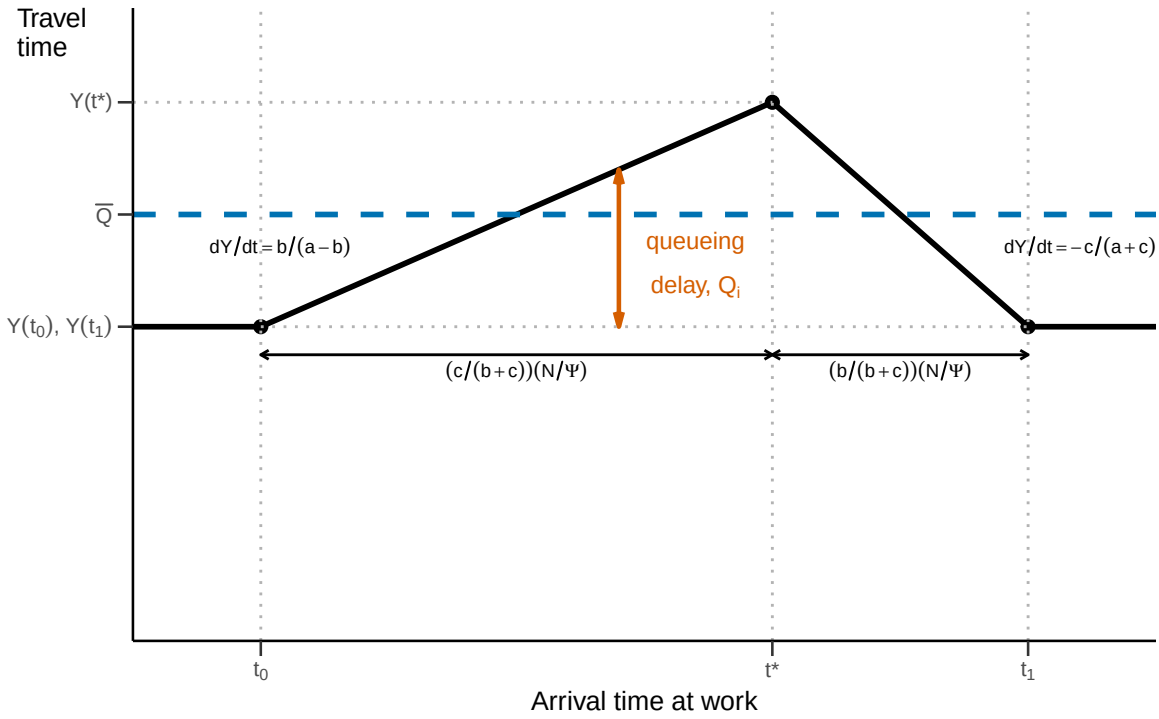
$$TQC = SDC = \frac{bc}{2(b+c)} \frac{N^2}{\psi}, \tag{16}$$

The total queueing delay is the aggregate time wasted sitting in the queue and is a pure deadweight loss. Each commuter ignores the extra delay her arrival imposes on those who follow. That unpriced marginal delay generates the queue. Dividing TQC by a and N yields the average queueing time:

$$\bar{Q} = \frac{b\,c}{2\,a\,(b+c)} \frac{N}{\psi}. \tag{17}$$

I denote the $bc/[a(b+c)]$ as a value-of-time composite, and N/ψ as the count-to-capacity ratio. The closed form makes clear that average queueing time scales proportionately with these parameters. Since a is a shadow price of time, it scales with wage (Small & Verhoef, 2007). Small (1982), and Kim and Moon (2024), showed that the ratio of schedule delay penalties b and c with a are statistically invariant across income groups. The composite parameter, and hence average queueing delay, should then be consistent across income groups.

In the homogeneous-users benchmark, total queuing and total schedule-delay costs are equal, and first-best time-varying tolls eliminate queues while leaving schedule delay unchanged (Arnott et al., 1993). Our empirical average queueing time recovers the deadweight triangle due to unpriced queuing.



Note: Adapted from Kim (2019).

Figure 2: Travel time as a function of arrival time.

3 Empirical framework

3.1 Main estimation model

The empirical strategy follows Kim's approach. Consider the ith commuter on a particular route with a bottleneck. The commuter's queueing delay is given by

$$Q_i \;=\; Y_i - Y_i(t_0) \tag{18}$$

as the difference between the observed travel time Y_i and the counterfactual free-flow travel time $Y_i(t_0)$ (equal to $Y_i(t_1) = x$), which is the travel time that would have been observed if the commuter had arrived early or late enough to avoid queueing. In Figure 2, Q_i is the vertical distance (orange) between travel time and the free-flow time. This is similar to the horizontal line queueing time between the cumulative arrival schedules in Figure 1, but the x-axis is now the arrival-at-work schedule.

From Figure 2, I can express the average queueing time on a single route as

$$\bar{Q} \;=\; \frac{\psi}{N}\int_{t_0}^{t_1}\left[Y(t)-Y(t_0)\right]dt. \tag{19}$$

Each arrival is time-weighted by the rate of arrival equivalent to the bottleneck capacity of ψ. To incorporate J heterogeneous routes (i.e., mutliple possible bottlenecks), I index each route by $j = 1, ..., J$, with N_j commuters, capacity ψ_j, and queue-forming window $[t_{0(j)}, t_{1(j)}]$. A function $f(j) \equiv N_j/\tilde{N}$ describes the distribution of commuters across routes, where $\tilde{N}$ is the total number of commuters across all routes. The overall mean queueing time is then just the weighted average of each route's integral in Equation 19, aggregating by $f(j)$:

$$\sum_{j=1}^{J} f(j)\,Q_j \;=\; \sum_{j=1}^{J}\frac{\psi_j}{\tilde{N}}\int_{t_{0j}}^{t_{1j}}\left[Y_j(t)-Y_j(t_{0j})\right]dt\;f(j). \tag{20}$$

At this point, I identify four empirical challenges to estimating $\bar{Q}$:

1. I do not observe arrivals at the bottleneck from my dataset, only departure-for- and arrival-at-work timings.
2. In the morning, each commuter only takes one trip to work.[5] I cannot observe both $Y_i(1)$ and $Y_i(0)$ for the same individual.
3. Each commuter can travel different origin–destination pairs, each with its own free-flow distance and capacity. These routes may contain unobserved attributes (speed limits, signal density, proximity to freeways etc).
4. Likewise, latent commuter preferences that may drive route choice and timing, such as risk tolerance, schedule flexibility or a taste for driving faster are not observed.

I address the unobserved arrival at the queue by using arrival-at-work time instead. On a given origin–destination, free-flow travel time is constant, so work-arrival time equals queue delay plus that fixed travel time. Departure-for-work time is a poor proxy for queue exposure due to the heterogeneity in trip distances. Someone leaving early on a long route may

[5]There are a very small minority of commuters in the dataset who reported multiple commute trips in the morning but it is unclear why that is the case. I exclude them from the main regression analysis but include them in a within-household fixed effects analysis.

still hit the queue, while a later departure on a short route may not. If I select commuters whose arrival-at-work fall within an empirically determined window, I can recover who entered the queue without observing queue-entry directly.

By defining a binary indicator based on arrival-at-work time, I compare the average travel time of those in the queue window to those who just missed it, holding constant all route and individual controls. Estimating this difference-in-means recovers the average treatment effect of queue exposure, and serves as my empirical estimate of the theoretical $\bar{Q}$.

To address observed selection, I control for commute distance and a wide array of neighbourhood, household and commuter covariates. Even with these covariates, there may still be unobserved route and commuter characteristics. For example, if a route is riddled with intersections where traffic signal timings changes during peak hours, commuters on that route may prefer to travel during off-peaks hours to avoid the delay from more frequent red lights. My estimation would then attribute their shorter off-peak travel times to the absence of queueing rather than to the route's inherent peak sluggishness. Thus, I explicitly insert unobserved route characteristics and timing preference terms into my estimation model.

The primary estimation model aims to recover the average queueing delay $\bar{Q}$ via a discrete-treatment regression that is similar to Kim's specification and identification strategy. I exploit three cross-sections (2001, 2009 and 2017) to sharpen identification and track how queueing delay evolves over time. In my baseline specification, I include survey-year dummy variables (2009 and 2017, with 2001 as the reference) and interact each with the queue-window indicator, allowing the average treatment effect of queue exposure to differ flexibly by wave. Year dummies also enter as standalone controls to soak up aggregate shifts in fuel prices, roadway investments, telecommuting uptake, and other broad trends. This structure lets me test whether bottlenecks' impact on travel time has intensified or abated between 2001 and 2017. I estimate,

$$Y_{ijk} = \alpha + \delta_k T_{ijk} + \sum_{p=1}^{3} \phi_p m_j^p + X_i'\gamma + u_j + w_{ij} + \lambda_k + \varepsilon_{ijk},$$

where

- Y_{ijk} is one-way commute time (minutes).

- $T_{ijk} = 1$ if arrival-at-work $\in [t_0, t_1]$, 0 otherwise. Cross-sectional survey years are represented by k.
- m_j is commute distance (tens of miles), with p polynomials representing the non-linear relationship between commute distance and commute time.
- X_i is a comprehensive set of commuter's neighbourhood, household and personal covariates. Neighbourhood variables include tract level population density and rental share,[6] block level dummies denoting the level of urban development, and MSA-level fixed effects, with dummies for each identifiable MSA, and the remaining unidentified MSAs collapsed into population-size bins. Household indicators include income and the number of workers. Individual indicators are age, gender and whether the individual carpools. There is also a day-of-week fixed effects to control for within-week commute variations.
- u_j and w_{ij} capture route characteristics and individual commuter unobservables.
- λ_k are survey year dummies (2001 omitted) that control for mean shifts across the cross-sections.
- ε_{ijk} is a mean-zero idiosyncratic error term.

With the average queue time by year, I can translate them into an aggregate monetary cost. To do so, I multiply the estimates by the total number of commutes within $[t0, t1]$ and an appropriate value of commute time, a. This converts my average bottleneck delay estimate into an aggregate welfare loss.

3.2 Alternative specifications

I assess the robustness of my estimated treatment effect to the explicit unobserved heterogeneity by estimating alternative specifications: household-fixed-effects (HHFE), instrumental-variable (IV) models, and interaction with the closed-form $\bar{Q}$ parameters.

The HHFE model purges time-invariant commuter preferences. IV models exploit exogenous predictors of arrival timing. I also interact my queue-exposure dummy with proxies of

[6] I use census tract instead of block group for these indicators since they are small enough to pick up reasonably meaningful neighbourhood variation while avoiding the sparseness and noise at the block-group level. Urban development dummies are only available at the block group level.

the closed-form $\bar{Q}$ parameters from the bottleneck model to investigate whether the treatment effect changes as theory predicts. These exercises jointly probe the consistency and structural validity of my main estimates.

The HHFE exploits variation in commute timing within multi-worker households to net out unobserved vehicle, residential location or route preferences. I include only controls that vary within household. By differencing out household effects, I isolate the impact of peak-time arrival on travel time free from time-invariant commuter traits. I also instrument peak arrival timing with individual job dummies. The underlying assumption made here is that job-driven rules like rigid shift-start or core hours schedules can induce exogenous shifts in work arrival timing without directly affecting route-specific travel time shocks. With valid instruments, this approach can recover local average treatment effects (LATE) for sub-populations defined by job categories. In my interaction models, I proxy the unobserved route-level parameters from Equation 17 to investigate whether the treatment effect changes as theory predicts. I also explore whether urban sprawl, aggregate road use intensity, toll roads and high-occupancy vehicle lanes have any observed changes to average queueing delay.

For MSAs with sufficient sample size ($n > 300$), I re-estimate the main estimation model yielding city specific queueing delay for 2017. I then use a log-linear version to estimate the proportional increase in mean commute time due to queueing. I interpret this as a city-level bottleneck indicator, Q-index, that allows me to benchmark my findings with the FHWA travel time index (Federal Highway Administration, 2018).

4 Data

4.1 Commute-level data

I use trip-level records from the publicly available National Household Travel Survey (NHTS) for 2001, 2009, and 2017. The NHTS employs a multistage design stratified by metropolitan statistical area (MSA) population groups to sample households, and collects a 24-hour trip diary from each household member. The diary provides trip purpose, mode, distance, departure and arrival times, and realised commute time, alongside household

and individual demographics. Trip-level survey weights allow nationally representative estimates and representation of major urbanised areas.[7]

I restrict the sample to weekday, one-way commutes by private-vehicle drivers arriving between 05:00 and 11:00. This window captures the morning peak and adjacent shoulders while preserving comparability across MSAs and survey waves. This time period accounts for about 98% of recorded commutes made between midnight to noon. I compute the survey-weighted estimates of total morning commute by summing the NHTS person–trip weights for my morning commute sample. This yields an estimated 15.86 billion commutes in 2001, 17.21 billion in 2009, and 18.2 billion in 2017.

In Section 2.1, the bottleneck window, $[t_0, t_1]$, is the period when commuter arrivals outpace roadway capacity and induce queueing delays. In principle, the true onset of queueing differs by link. While bottlenecks across routes can start at different time, I follow Kim's approach of choosing a single common threshold such that at least one commuter on every route travels in free-flow before the threshold.

Figure 3a compares mean travel times in 30-min blocks over the morning commute sample. The first cut-off with a statistically significant positive difference-in-means is at 6 AM. I select this as t_0 and let t_1 be 11 AM to allow the capture of all sustained queueing effect. I trace queueing-time profile within this interval via bin-dummy regressions in Section 5 to test the threshold assumption.

Figure 3b plot log travel-time-per-mile against distance for a random 10% subsample of morning commutes across the survey years. The linear fit shows the standard decline in time-per-mile with distance; the fifth-order polynomial shows mild curvature. Following Kim, I include distance polynomials up to the fifth degree in the main specification to allow flexible elasticities rather than imposing a single log-distance control.

Table 1 shows the summary statistics of variables form the NHTS data that are used in my analysis. To recap, commute time (Y) is the dependant variable in my specification. The key regressor is the bottleneck-window dummy (arrive 6-11 AM, $T = 1$), which identifies congestion exposure relative to shoulder/off-peak trips on the same routes. The specific trip-level controls are commute distance and day-of-week. Neigbourhood or implicit

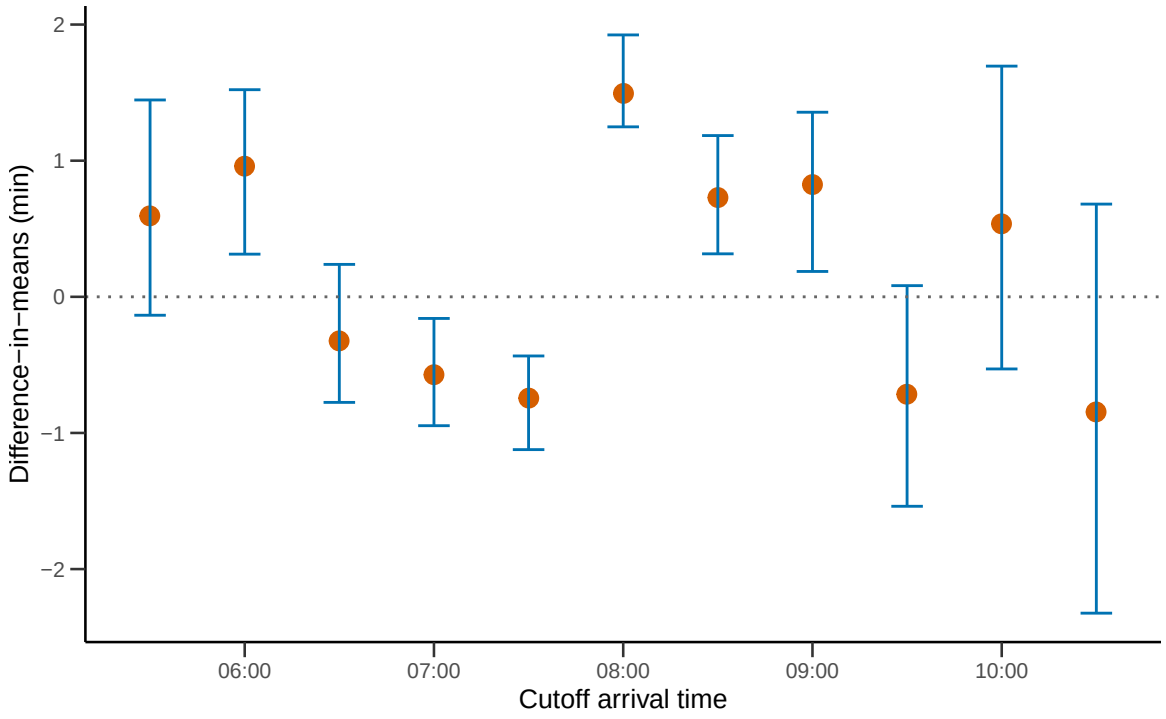
[7]The NHTS changed its survey methodology after 2017. The 2022 data is only representative down to the census division without tract level household residential data. It is incompatible with my MSA-level estimation approach.

trip controls include tract populaiton density, rental profile, level of urban development, and MSA-level fixed effects. For individual and household controls, we include age, gender, race, household income, workers per household, vehicle ownership, and household-with-children dummy. These absorb compositional differences correlated with travel time. Industry/occupation dummies are used as proxy for start-time rigidity.

Average morning commute time has risen from 2001 to 2017 by about 19% while the average commute distance held steady at about 12 to 13 miles. The summary data also depicts a pattern of urban concentration over the study period. The share of commuters living in MSAs with more than one million people grew from 37% in 2001 to 49% in 2009, before settling at 46% in 2017. Tract population density also climbed, reflecting denser residential locations. Commuters with professional jobs rose from 45% to 57% of the sample, while clerical, manufacturing and service role shares declined. These patterns suggest that heavier congestion and work schedules, not longer routes, drive the increasing travel times.

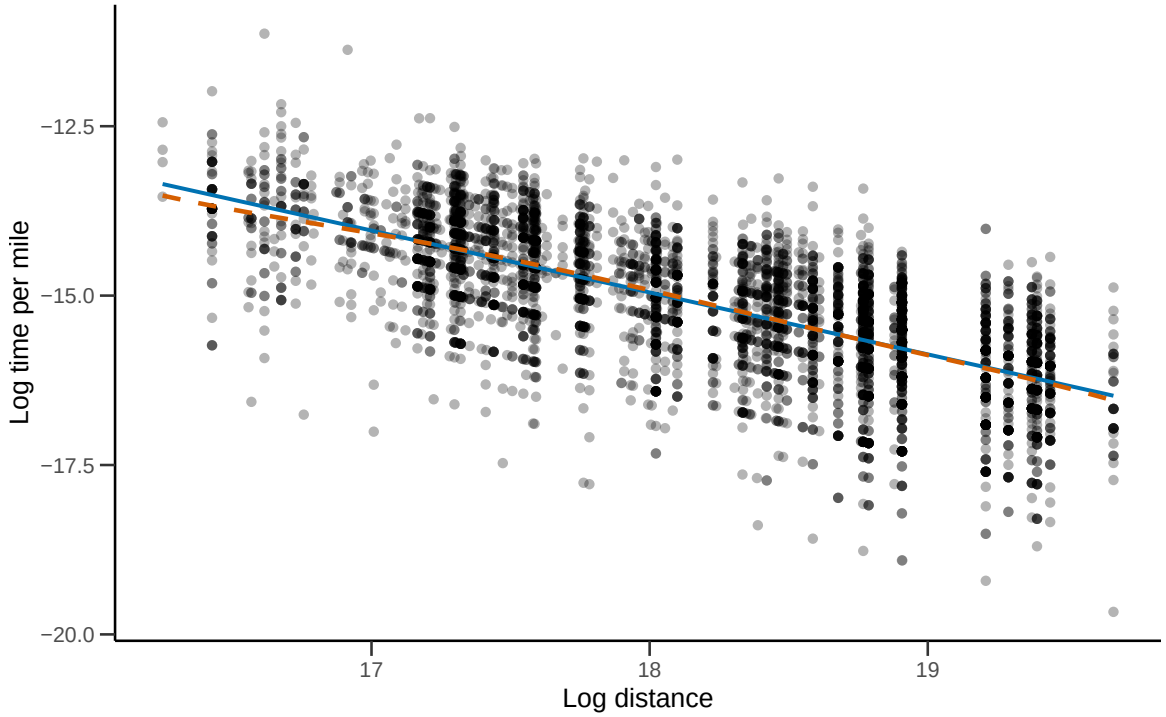
The NHTS public releases only provide unique identifiers to select MSAs with more than 1 million population. Commutes from these MSAs accounted for 56% of the morning commute sample. This ratio remains about the same when comparing commutes within the chosen bottleneck window. Since my study span 2001 to 2017, MSAs and their component set of counties can change across waves. For example, Los Angeles-Riverside-Orange County (id: 4472) in 2001 became Los Angeles-Long Beach-Santa Ana (id: 31100) in 2009, and Los Angeles-Long Beach-Anaheim (id: 31080) in 2017. I account for these changes in MSAs, identifers and county components in building the MSA-level data (see Section 4.2).

I estimate the main model on both the full NHTS sample and the identified MSA subsample under the assumption that persistent, capacity-driven bottlenecks occur predominantly within larger metropolitan areas. My full-sample estimates capture the average queueing time across all private vehicle commute types nationally, reflecting both recurring and non-recurring delays. The MSA-only estimates, by contrast, concentrate on large metro areas where route geometry is more uniform, free-flow times are measured more reliably, and true peak-period queues can prevail. Comparing these two sets of results lets me test whether the average treatment effect of queue exposure generalizes nationally or is driven primarily by urban bottleneck dynamics.



Note: Error bars show 95% CI.

(a) Difference-in-means of commute time by arrival-time cutoff



Linear fit (solid blue), 5th−order poly fit (dashed orange)

(b) Commute time and commute distance

Figure 3: Difference-in-means by arrival-time cutoff, commute time and commute distance.

Table 1: Summary statistics for NHTS data

	2001		2009		2017	
	Mean	SD	Mean	SD	Mean	SD
Commute time (mins)	21.2	(17.3)	23.2	(18.3)	25.2	(19.5)
Arrived between 6 AM and 11 AM	0.90	—	0.92	—	0.91	—
Arrival time at work (mins from 5 AM)	165.1	(76.1)	171.8	(75.5)	177.4	(77.9)
Commute distance ('0 miles)	1.2	(1.5)	1.3	(1.4)	1.3	(1.4)
Tract population density ('000 per sq mile)	27.9	(43)	27.9	(39.4)	33.1	(43)
Tract rental units (percent)	26.1	(18.9)	25.1	(18.2)	27.2	(19.1)
Residence in MSA with >1M+ population	0.37	—	0.49	—	0.46	—
Residence in MSA with <1M population	0.46	—	0.31	—	0.41	—
Neigbourhood category is urban	0.06	—	0.08	—	0.11	—
Neigbourhood category is second city	0.20	—	0.18	—	0.20	—
Neigbourhood category is suburban	0.23	—	0.25	—	0.25	—
Real household income ('000 2017 USD)	88.5	(43.7)	89.3	(39.5)	99.3	(58.8)
Household size	3.0	(1.3)	2.9	(1.3)	2.7	(1.3)
Workers per household	2.0	(0.8)	1.8	(0.7)	1.8	(0.7)
Home ownership	0.84	—	0.92	—	0.80	—
Vehicles per household	2.5	(1.2)	2.6	(1.2)	2.5	(1.2)
Household with children	0.42	—	0.34	—	0.30	—
Race is white	0.89	—	0.87	—	0.84	—
Female	0.47	—	0.49	—	0.49	—
Age	42.7	(11.7)	48.4	(12)	46.4	(13.5)
Clerical job	0.14	—	0.15	—	0.12	—
Manufacturing job	0.20	—	0.14	—	0.13	—
Professional job	0.45	—	0.50	—	0.57	—
Service job	0.21	—	0.21	—	0.18	—
Observations	22,711	—	45,182	—	47,948	—
Commutes 5-11am ('000,000)[1]	15,855	—	17,209	—	18,199	—
Commutes 6-11am ('000,000)[1]	14,070	—	15,330	—	16,156	—
Identified MSAs with >1M+ population	50	—	51	—	52	—
Commutes 5-11am, MSAs ('000,000)[1]	9,046	—	9,179	—	10,350	—
Commutes 6-11am, MSAs ('000,000)[1]	8,110	—	8,227	—	9,258	—

[1]Author's calculations using NHTS survey weights.

4.2 Metropolitan-level data

I pair the NHTS dataset with road capacity data from the Federal Highway Administration (2016) Highway Performance Monitoring System (HPMS). The HPMS reports yearly county-level total lane-miles of interstate highways (FWY) and major urban roads (MRU), annual average daily traffic (AADT) for the road segments, and the presence of toll and hig-occupancy vehicle (HOV) lanes. I retrieve aggregate vehicle miles travelled (VMT) by multiplying AADT with road segment lengths. These parameters are then aggregated up to the MSA-level for MSAs that are identified in the NHTS.

Table 2: Summary statistics at MSA-level

	2001		2009		2017	
	Mean	SD	Mean	SD	Mean	SD
Daily VMT ('000,000 miles)	56.1	(59.1)	61.0	(50.5)	67.4	(55.1)
Interstate highways, FWY ('000 lane-miles)[1]	1.4	(1.2)	1.4	(0.8)	1.5	(0.9)
Major urban roads, MRU ('000 lane-miles)[2]	7.1	(7.2)	8.2	(6.5)	8.5	(6.8)
Toll segments (miles)	35.0	(61.6)	43.2	(68.7)	55.8	(85)
HOV segments (miles)	11.5	(22.6)	18.8	(37.3)	26.1	(45.7)
MSA population ('000,000)	3.2	(3.8)	3.2	(3.2)	3.4	(3.3)
Commuters by car, truck or van ('000,000)	1.3	(1.2)	1.3	(1)	1.4	(1.1)
Real weekly wages ('000 2017 USD)	1.0	(0.1)	1.0	(0.2)	1.1	(0.3)
Identified MSAs with >1M+ population	50	—	51	—	52	—

[1]All interstate highway segments within MSA are included.
[2]Major urban roads include non-interstate highways, principal arterials, minor arterials, and collectors. Local roads are excluded due to inconsistent reporting.

Over the study period, the HPMS public releases transitioned from tabulated to geospatial formats (Federal Highway Administration, 2008b), which led to some attribute fields being added, dropped, or restructured. To ensure comparability, I harmonise these inconsistencies under the assumption that physical roadstock remains stable over short intervals. For example, I used the 2011 HPMS release to align with the 2009 NHTS since no public HPMS data is available for 2010 and data was truncated from 2007 to 2009. The 2017 geospatial dataset does not include HOV-lane flags, and the 2016 release lacks county identifiers needed for counties-to-MSA crosswalk. To restore consistency, I extract HOV information from the 2015 HPMS and merge it into the 2017 file. Likewise, because counts of toll and HOV lanes themselves are not reported uniformly across years, I standardise on total miles as the indicator of toll and HOV capacity.

I used other sources such as the U.S. Census, via Manson et al. (2024), and the U.S. Bureau of Labour Statistics (BLS) Current Employment Statistics, CES (Bureau of Labor Statistics, 2017), and Quarterly Census of Employment Statistics, QCEW (Bureau of Labor Statistics, 2025), to construct interaction variables. For example, I use QCEW wage data at the MSA-level as candidates to proxy schedule unit costs $[a, b, c]$ in the interaction models. Table 2 shows the summary statistics for a collection of main MSA-level data.

In the next section, I first recover the average queueing delay using the main specification, then explore its robustness using alternative specifications, closed-form $\bar{Q}$ proxies, and other possible determinants of bottleneck congestion. I will also present average queueing delay for selected MSAs.

5 Results

5.1 Bottleneck queueing delay

Table 3 reports the primary regression estimates of average queueing time. In Column (1), drivers arriving within the 6 to 11 am window commute 2.84 minutes longer on average across years. Column (2) shows the change in average queueing delay from 2001 (1.76 minutes) to 2017 (3.57 minutes). Adding neighbourhood, household/commuter variables and fixed effects in Columns (2) to (5) cuts the estimated queue penalty to 1.45 minutes in 2001, 2.44 minutes in 2009, and 3.25 minutes in 2017. I interpret this as a sharpening of identification by comparing commuters on similar routes. In other words, once variation in route and individual characteristics is netted out, the estimations reflects bottleneck delay rather than differences in routes or neighbourhoods.

I compared the sample unweighted and weighted treated effects for the model in Column (5) using a joint Dumouchel-Duncan test (Dumouchel & Duncan, 1983). I fail to reject the null ($H = 4.96, p = 0.174$) that sampling weights materially alter the estimates. Therefore, I only report unweighted results. Although the 2022 NHTS lacks neighbourhood and MSA variables used in the main analysis, I report 2022 estimates with Census Division controls and fixed effects in Appendix A as a robustness check.

The estimates from Column (5) suggest that residual equilibrium bottleneck queues persist, and even grow, across the years. Under the bottleneck model, this can imply that peak-period commute demand continues to be relatively inelastic with respect to congestion and outpace bottleneck capacity expansion. Capacity expansions has been shown to induce proportional increases in vehicle-miles travelled (Duranton & Turner, 2011). This may also contribute to the increasing queueing delay during the period of study.

Column (6) restricts the sample to identified MSAs with more than 1 million population. Here, the estimated penalty rises to 3.14 minutes in 2001, 4.17 minutes in 2009, and 5.57 minutes in 2017. This aligns with the bottleneck model and the closed-form $\bar{Q}$ function, which predict that average queueing delay scales positively with commuter mass and inversely with capacity. Larger MSAs, with their higher population density, concentrated employment centres, and converging arterial networks, can also funnel commuters into the same constrained corridors with limited viable detour options (Geroliminis & Daganzo, 2008; Needell et al., 2023). Such structural characteristics can amplify queue formation and

homogenize delay across household members, driving up the average bottleneck penalty. The disproportionately larger growth outside major metros suggests that emerging bottlenecks in smaller or lower-density regions are outpacing network expansions, perhaps driven by rising vehicle ownership, changing job rules, and uneven roadway investment.

With limited scheduling flexibility outside of the bottleneck window due to commute demand inelasticity, commuters under the bottleneck regime are expected to adjust arrival rates until their queueing delay cost equals schedule delay cost (see Equation 5). Figure 4 plots estimated queueing delay by arrival-at-work intervals, both for the pooled sample and separately in 2001, 2009 and 2017. All profiles generally start close to zero before 6:30am, validating the threshold chosen in Section 4.1, and peak between 8:30am to 9:30am, reflecting the rising part of the queue profile in the bottleneck model.

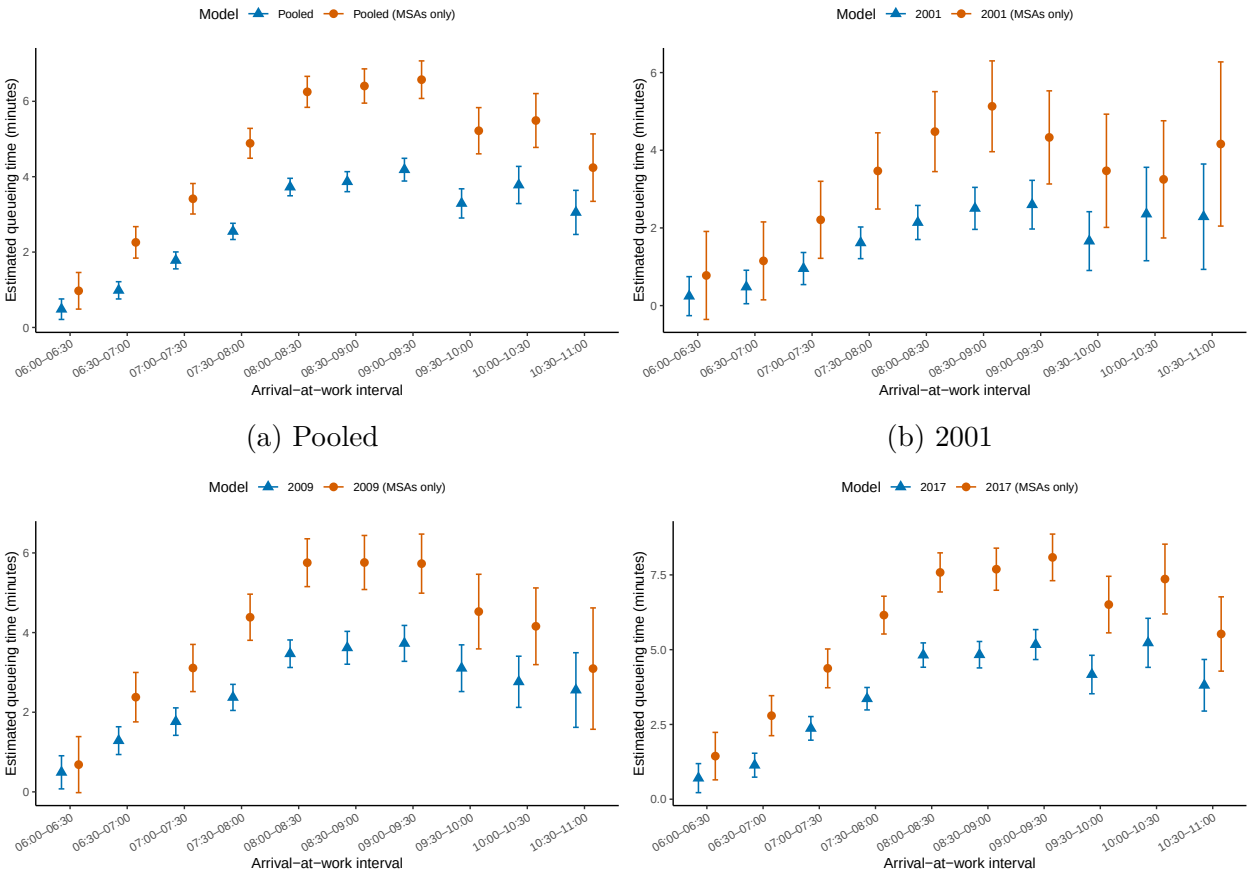
While queueing time is expected to fall after peaking like the 2009 profile (Figure 4c), that is not the case for the other profiles where a plateauing trend is observed instead, with greater variance. Kim offers a few possible explanations for this. A rising share of trips after 10:00am can be for purposes other than the standard home-work commute. Trips for business, shopping or leisure can have more dispersed origins and destinations, so they tend to re-enter bottlenecks at odd times and inflate the estimated delay. The share of these non-commute trips can also be heterogeneous across areas, and can be the cause of the larger late-peak variance in the larger MSAs seen in the profiles. Finally, Kim also suggest that as the morning peak wears on, there is more opportunity for crashes, breakdowns and other traffic incidents to occur, injecting random blockages and sporadic congestion spikes.

Hall (2024) provides an empirical explanation by structurally embedding inframarginal commuters into the bottleneck model. Inframarginal commuters do not face a binding marginal trade-off between travel time and schedule delay. Hall shows that only the marginal subset of commuters actually equalise queueing delay with schedule-delay cost. The rest simply bunch at their desired times, flattening observed queueing profiles. The queue-time profiles in Figure 4 line up with Hall's inframarginal equilibrium rather than the standard knife-edge spike under homogeneous preferences shown in Figure 2. This interpretation is further supported by Kim and Moon (2024) who estimated a marginal rate of substitution between travel time and schedule delay in California. They found that this value is more than that of the majority of commuter travel-time profile slope, implying that the majority of Californian commuters are inframarginal and do not adjust departure times for small time-savings.

Table 3: Estimates of average queueing time within bottlenecks.

	Commute time					
	(1)	(2)	(3)	(4)	(5)	(6)
Bottleneck window	2.84***	1.76***	1.52***	1.55***	1.45***	3.14***
	(0.098)	(0.191)	(0.190)	(0.191)	(0.186)	(0.468)
Bottleneck window × 2009		0.759***	0.882***	0.949***	0.992***	1.03*
		(0.242)	(0.242)	(0.242)	(0.239)	(0.535)
Bottleneck window × 2017		1.81***	1.77***	1.82***	1.80***	2.43***
		(0.253)	(0.254)	(0.254)	(0.250)	(0.540)
Distance variables	✓	✓	✓	✓	✓	✓
Neigbourhood variables			✓	✓	✓	✓
Household/commuter variables				✓	✓	✓
Observations	112,745	112,745	112,745	112,745	112,745	49,150
R^2	0.728	0.733	0.743	0.745	0.750	0.708
Day-of-week fixed effects					✓	✓
MSA fixed effects					✓	✓

OLS regressions with a constant from (1) to (4). OLS fixed effects for (5) and (6). Observations in (6) are for identified MSAs in NHTS. Heteroskedasticity-robust standard errors in parentheses. $***p < 0.01, **p < 0.05, *p < 0.1$.



(a) Pooled

(b) 2001

(c) 2009

(d) 2017

Figure 4: Estimated queueing time by arrival-at-work intervals (error bars, 95% CI).

5.2 Within households and across jobs

Table 4 presents robustness checks for the main average queueing-time estimates. Unlike the OLS sample, which selects only respondents with one commute per day, the HHFE sample uses every observed commute within multi-worker households to isolate within-household variation. The 2001 within-household estimates in Column (1) of Panels A and B (1.22 min, 3.31 min) closely match its OLS analogue in Table 3 Columns (5) and (6) (1.45 min, 3.14 min). The interaction terms for 2009 and 2017 are both statistically insignificant, reflecting the lack of within-household variations across survey waves to exploit. The modest adjustment from OLS to HHFE suggests that unobserved, time-invariant commuter characteristics are unlikely to bias the primary estimates.

In Column (2) of Panel A, job-category covariates are added to the main model (Column 5 of Table 3). The estimated bottleneck-window effect falls only slightly from 1.51 min to 1.41 min. This stability supports the assumption that job categories do not absorb unobserved route-level travel time shocks.

The results of using job categories as instruments are shown in Columns 3 to 6. For the full sample in Panel A, the initial estimates for manufacturing (4.09 min), professional (5.81 min) and clerical (9.12 mins) jobs exceed the OLS estimate while service remains imprecise. Restricting to the identified MSAs sample in Panel B increases these effects. Notably, the 2017 $\times$ treatment interaction turns negative for professionals in Panel A and falls sharply for manufacturing in Panel B.

I examine the validity of the IV estimations by considering each job sector. Manufacturing jobs are typically organised around rigid, multi-shift production schedules so start times are essentially exogenous to traffic conditions. This shift-work structure injects ample variation in who falls inside the 6 to 11am window. Professional jobs tend to operate under core-hours policies but allow for flexibility to start times, allowing for variability in and out of the window. In contrast, clerical staff generally face standard fixed office hours with arrival times within the bottleneck window. Service-sector roles also face similar constraints through fixed business-opening times. As such, I expect LATE of manufacturing and professional jobs to reflect the queueing penalty borne by those in the bottleneck window but not that of clerical and service jobs.

In many large MSAs, the 2000s and 2010s saw the rollout of HOV or HOT lanes on key

Table 4: Estimates of average queueing time using alternative specifications.

Panel A	Commute time					
	OLS		2SLS			
	(1)	(2)	(3)	(4)	(5)	(6)
	HHFE	Jobs	Manufacturing	Professional	Service	Clerical
Bottleneck window	1.22**	1.41***	4.09***	5.81***	-8.20	9.12***
	(0.604)	(0.188)	(1.07)	(2.06)	(11.7)	(3.00)
Bottleneck window x 2009	1.12	0.994***	-0.042	-1.41	83.0	5.03
	(0.770)	(0.239)	(1.40)	(2.58)	(73.1)	(4.04)
Bottleneck window x 2017	1.27*	1.80***	-2.33	-7.24***	-47.2	-2.63
	(0.764)	(0.251)	(1.50)	(2.62)	(79.0)	(4.53)
Observations	119,807	112,745	112,745	112,745	112,745	112,745
R^2	0.953	0.750	0.750	0.748	0.178	0.738

Panel B	Commute time					
	OLS		2SLS			
	(1)	(2)	(3)	(4)	(5)	(6)
	HHFE	Jobs	Manufacturing	Professional	Service	Clerical
Bottleneck window	3.31**	3.02***	13.9***	20.5***	62.6	19.4**
	(1.41)	(0.469)	(3.48)	(7.30)	(213.5)	(9.60)
Bottleneck window x 2009	0.477	1.02*	-4.12	-10.6	-179.1	18.9*
	(1.67)	(0.535)	(3.76)	(7.77)	(533.3)	(11.1)
Bottleneck window x 2017	0.895	2.42***	-8.00**	-18.6**	-231.8	1.22
	(1.63)	(0.541)	(3.84)	(7.62)	(417.8)	(12.2)
Observations	51,713	49,150	49,150	49,150	49,150	49,150
R^2	0.947	0.708	0.703	0.698	-1.88	0.623

Full sample for panel A and only identified MSAs for panel b. All models include the same control variables and fixed effects as in column (5), Table 3. The instrument variable for 2SLS is listed below each column number. 2SLS 1st stage F-stats > 30 except (5), both panels, and (6), panel B; all Wu-Hausman test $p < 0.05$. Heteroskedasticity-robust standard errors in parentheses. $***p < 0.01, **p < 0.05, *p < 0.1$.

commuter corridors serving high-skill and industrial employment centres (Federal Highway Administration, 2009, 2024). By 2017, a subset of professional and manufacturing compliers whose start-work time are fixed by job rules may have gained preferential access to these lanes. Thus, it is plausible that the measured commute in the 6–11 AM window for 2017 is lower the marginal off-peak trip. Improvements to direct off-ramps, dedicated truck lanes or shuttle services in industrial hubs may have the same effect. I also suspect that that increasing core-hour flexibility and telecommuting trends may have caused the shift for professionals, as a smaller group is now isolated whose exogenous fixed peak-period arrival actually avoid the worst of the general-population queueing.

Together, these factors suggest that targeted interventions like lane pricing, staggered start incentives, or employer-level flex-scheduling can deliver the greatest travel-time relief to those sectors where scheduling rules admit genuine timing choice in and out of the bottleneck window. In contrast, sectors with near-zero timing discretion show no identifiable LATE and may stand to gain little from these interventions alone without consideration for the bottleneck window.

5.3 Determinants of bottleneck

The main regression is designed to estimate the average queueing delay $\bar{Q}$ which theory (Equation 17) indicates will scale proportionally with the count-to-capacity ratio, N/ψ and the value-of-time composite, $bc/a(b+c)$ (see Section 2.2). Since I lack direct measures of these structural parameters, I test year-based MSA-level proxies and covariates with the main regression model by interacting them with the bottleneck window dummy. Year interactions with the bottleneck window dummy and MSA fixed effects are dropped.

I use the daily number of person-trips in the 6–11 AM window, aggregated from NHTS person-trip weights, to proxy for N. To capture ψ, I draw from the HPMS data two measures of roadstock, FWY lane-miles and MRU lane-miles. Since the count-to-capacity ratio defines the presence of bottleneck congestion, I include their proxies in all estimations. In Appendix A, a robustness check is done using the census-derived number of commuters from MSAs instead of trips from the NHTS.

I employ three complementary wage-based proxies, taken from BLS QCEW, for the composite parameter. First, I use the average weekly wage (AWW), which most directly re-

flects the per-unit-time value of time and schedule delay on an everyday commute.[8] Second, I include average annual earnings (AAW) to gauge how longer-run, across-MSA income differences map onto daily trade-offs in commuting. Finally, I use total MSA payroll (TAW) as an aggregate indicator that bundles both the scale of wages and the size of the commuting workforce. I drop commuter count for total payroll to avoid collinearity.

To capture broader built-environment characteristics and their impacts on bottleneck congestion, I include the overall and component sprawl indices of Hamidi et al. (2015) and Ewing et al. (2003). They define sprawl as occurring when the pace of development significantly exceeds that of population growth. The indices score each MSA on concentration of activity, land-use mix, development density, and street-network connectivity against a national baseline of 100, with scores lower than 100 indicating more sprawl. Since the period of study for these indices only correspond to the 2001 and 2009 waves, I drop the 2017 observations for estimations with these indices.

I also include total VMT, toll road capacity, and HOV lane capacity as standalone covariates. Total VMT measures road usage, and together with the total roadstock, captures realised demand intensity. Additional tolled capacity, measured in miles, may reduce volume, or simply reroute traffic, so including them tests an extended capacity effect (also see the section on bottleneck pricing, Section 6.2). Similarly, HOV capacity (also in miles) can affect peak flow and incentivise ridesharing, moderating both congestion and scheduling choices. Together, these variables can be used to assess the robustness of the closed-form bottleneck intuition in explaining observed variations in average queueing delay across the MSAs.

Table 5 reports estimates of each proxy and covariate interacted with the 6–11 AM bottleneck dummy. Since the dependent variable is commute time in minutes, coefficients on log-transformed regressors are semi-elasticities (minutes per log-point), while coefficients on level variables (like sprawl) are marginal effects (minutes per unit). For the structural proxies, only commute counts consistently behave as the closed-form predicts, with larger observed demand pushing up commute delays. I attribute this to the survey-weight sum, despite aggregation, being a reasonable approximation of true instantaneous flow at bottlenecks. By contrast, the proxy FWY-MRU measure is a broad time-invariant stock compared to the true ψ which measures count per unit time. The route-specific controls in the

[8] BLS QCEW 2017 does not include average weekly wages data. Not all of the identified MSAs have 2017 hourly wage data from other sources.

Table 5: Estimates of proxies and covariates of bottleneck congestion.

	Commute time			
	(1)	(2)	(3)	(4)
Panel A	Proxies	log(AWW)	log(AAW)	log(TAW)
log(Commutes 6-11am)	1.67***	0.673	0.673	
	(0.601)	(0.659)	(0.659)	
log(FWY Lane Miles)	-0.609	-0.273	-0.273	-0.339
	(0.513)	(0.514)	(0.514)	(0.493)
log(MRU Lane Miles)	0.150	0.225	0.225	0.207
	(0.616)	(0.626)	(0.626)	(0.508)
Added variable		5.58***	5.58***	1.46***
		(1.42)	(1.42)	(0.356)
Observations	49,150	49,150	49,150	49,150
R^2	0.704	0.705	0.705	0.704

	Commute time			
	(1)	(2)	(3)	(4)
Panel B	Sprawl	Centeredness	Use Mix	Density
log(Commutes 6-11am)	0.282	0.483	0.261	0.281
	(0.675)	(0.657)	(0.680)	(0.686)
log(FWY Lane Miles)	0.483	0.283	0.381	0.285
	(0.657)	(0.625)	(0.628)	(0.619)
log(MRU Lane Miles)	0.748	0.917	0.585	0.549
	(0.718)	(0.716)	(0.725)	(0.732)
Added variable	0.016	0.017	0.028**	0.023**
	(0.010)	(0.014)	(0.011)	(0.011)
Observations	27,700	27,700	27,700	27,700
R^2	0.730	0.730	0.730	0.730

	Commute time			
	(1)	(2)	(3)	(4)
Panel C	Network	log(VMT)	log(Toll)	log(HOV)
log(Commutes 6-11am)	0.418	0.123	1.72***	1.61***
	(0.650)	(0.671)	(0.595)	(0.597)
log(FWY Lane Miles)	0.629	-2.01***	-1.06**	-0.517
	(0.644)	(0.577)	(0.533)	(0.515)
log(MRU Lane Miles)	0.245	-3.11***	0.849	-0.126
	(0.777)	(0.878)	(0.656)	(0.655)
Added variable	0.020*	5.80***	-0.256***	0.187*
	(0.010)	(1.03)	(0.093)	(0.107)
Observations	27,700	49,150	49,150	49,150
R^2	0.730	0.705	0.705	0.705

Only interaction terms reported. Identifed MSA sample for all models. 2017 excluded for B1-4, C1. Same controls as column (5), Table 3, without MSA fixed effects. Heteroskedasticity-robust standard errors in parentheses. $***p < 0.01, **p < 0.05, *p < 0.1$.

original model will have already soaked up the cross-MSA variation in effective bottleneck capacity. Thus, this leaves little residual variation for the lane-mile proxy to register a significant effect.

When the three wage-based proxies are introduced, the story becomes more nuanced. As discussed in Section 2.2, one might expect that the composite value-of-time parameter to be invariant across income. However, all three wage-based proxies enter positively and significantly (Panel A, Columns 2 to 4), absorbing much of the cross-MSA variation of the proxies. I interpret this in two ways. The wage measures may be picking up variations in correlated unobserved factors like commute mode accessibility and telecommuting rates (Dingel & Neiman, 2020; Zhu, 2013). Second, taking into account the changes in the N and ψ proxy estimates, an increase in wage is moderating the slope of the proxy count-to-capacity ratio akin to the inframarginal traveller effect described by Hall (2024) (see Section 5.2). Since wages alone do not fully capture the composite value-of-time trade-off (i.e., endogenous scheduling), empirical measurement of the structural parameters $[a, b, c]$ remains crucial for structural bottleneck analysis.

Based on the sprawl estimate from Panel B (1) to (4) and Panel C (1), more compact and better-connected metro areas exhibit slightly longer 6 to 11am commute times holding commute count and capacity constant. Compact MSAs tend to concentrate jobs, housing, and traffic onto fewer key corridors (Glaeser et al., 2001), which is exactly the same effect that, mechanically, a commute demand or lower capacity would produce in the structural model. However, more compact grids may also give commuters many alternative routes to ease pressure on any single link (Ewing & Cervero, 2010). Only land-use mix and use-density are significant but the effects of the proxies are generally absorbed. I interpret this as the built-environment and neighbourhood controls in the main specification exerting a measurable effect.

The elasticity of commute time with respect to VMT is 5.8 (Panel C, Column 2) which is much larger compared to the commute time with respect to commute count (1.67) in the proxy-only model (Panel A, Column 1). This finding accords with VMT being widely accepted in the transport literature as a direct measure of aggregate demand since it captures both trip frequency and distance travelled. Duranton and Turner (2011) further show that changes in VMT closely track demand shifts when evaluating induced-travel effects of added highway capacity in U.S. urban areas. I posit that this is why the general FWY-MRU roadstock proxies now track significantly with queueing delay. While the main spec-

ification relies on discrete commuter counts (since individual departure decisions are endogenous), future work could enrich the bottleneck framework by considering VMT as a continuous demand feature.

In Panel C (3) and (4), additional tolled capacity correlates to a drop in peak commute time by 0.256 minutes per log-point, whereas adding HOV capacity correlates to an increase of 0.187 minutes. Theoretically, time-varying tolls can fully eliminate queueing delays by charging drivers the marginal social cost of waiting, as first shown by Vickrey and formalised in the bottleneck framework by ADL (see Section 6.2). In practice, however, most toll facilities use coarser pricing rules (e.g., flat or peak/off-peak rates), which can only partially clear the queue if commute demand within the peak-period is elastic. Thus, the toll estimate may be reflecting this moderation.

The positive correlation of HOV capacity with queueing delay can seem counter-intuitive but there are two straightforward interpretations consistent with the bottleneck framework. First, if the additional HOV lanes do not substantially increase overall corridor capacity, then their opening shifts the binding bottleneck onto the remaining non-tolled lanes, leaving total throughput unchanged or even worse off (Daganzo & Cassidy, 2008). Second, even when overall capacity is expanded, these exclusive lanes may attract only a subset of drivers as suggested earlier in Section 5.2. Both channels effectively raise the demand-to-capacity ratio faced by non-HOV commuters, prolonging wait times. While I corroborate my estimates with theoretical mechanisms, I are unaware of any recent U.S. MSA-level empirical work establishing a causal link between actual toll and HOV capacity, and bottleneck delays.[9] This is an important avenue for future research.

5.4 City-level bottleneck

Table 6 summarises the estimates of average queueing delay in 2017 for the U.S. as a whole, identified MSAs, and then MSAs (named by their core city) with more than 300 observations. The first two columns show the proxy for count-to-capacity ratio from Section 5.3, using weighted trip counts and road stock, and the number of recorded observations

[9] Cohen et al. (2022) provide insights into how HOV lanes increase carpooling using a natural experiment in Israel. Brownstone and Small (2005) review 5 stated-preference and revealed-preference dataset from two high occupancy toll (HOT) lane implementations in South California. Brent and Gross (2018) and Goel and Burris (2012) document HOT lane implementations, pricing, and responses from users. HOT are a mixture of toll and HOV lanes where vehicles are charged for using the lane.

n. The third column shows the average queueing delays, $\bar{Q}$, in minutes, as estimated via the main model (Table 3, Columns 5 and 6), followed by the 95% confidence interval. The next column, Q (log), reports the same effect measured as a semi-elasticity (i.e., the log of commute time replaces commute time in minutes as the dependant variable). Exponentiating the semi-elasticity yields the multiplicative "Q-Index", which is the factor by which average commute time is inflated by bottleneck queueing delay alone. Finally, TTI gives the average of the quarterly Travel Time Indices (TTIs) reported by FHWA in their urban congestion reports. This index is the ratio of peak-period travel time to free-flow travel time, and is only published at the MSA level. I provide the city-level measures for 2001 and 2009 in Appendix A.

The city queueing delay estimates range from just 1.81 min in Charlotte up to 11.8 min in San Jose. The averaged estimate for all 15 MSAs is 6.55 min, which is larger than the 2017 national (3.25 min) and all identified MSAs (5.57 min) estimates. Bottleneck delays alone typically add 12–36 percent to metropolitan one-way morning commute times with most cities exhibiting a Q-Index between 1.12 and 1.36, with the identified MSAs Q-index being 1.2. The largest queue-driven inflation is in San Jose (1.63), followed by Los Angeles (1.36), Chicago (1.34), and Austin (1.27). Charlotte (1.01) and Houston (1.12) lie at the lower end. The national level Q-index falls on the lower end of the scale at 1.12.

Unlike the analysis in this study which relies on travel-survey data to estimate pure bottleneck delay, the FHWA's Travel Time Index is calculated from passive-probe vehicle data that records actual speeds, trip durations, and roadway volumes. A long-standing counterpart index reported in the Texas A&M Transportation Institute Urban Mobility Report (Schrank et al., 2019) uses the same probe-based methodology (and mostly the same data sources), so I compare only the FHWA metric here for brevity. Although the Q-Index and the FHWA TTI generally rank cities in the same order, the Q-Index values are consistently lower. That is because the Q-Index isolates only the pure bottleneck delay—holding free-flow speeds and off-peak departure patterns fixed whereas the FHWA TTI measures the ratio of actual peak-period travel time to free-flow travel time on the entire urban network, capturing all forms of slowdown (e.g., arterial signals, on- and off-ramp merging, and non-bottleneck queueing).

Table 6: Bottleneck congestion by cities in 2017.

	N/ψ[1]	n[2]	$\bar{Q}$ (mins)[3]	95% CI	Q (log)[4]	95% CI	Q-Index[5]	TTI[6]
USA[7]	41.2	112,745	3.25	[2.92, 3.58]	0.11	[0.10, 0.13]	1.12	-
MSAs[7]	49.8	49,150	5.57	[5.03, 6.11]	0.18	[0.16, 0.20]	1.20	1.33
Atlanta	54.4	1,212	7.84	[4.93, 10.75]	0.13	[0.03, 0.23]	1.14	1.37
Austin	63.1	992	7.07	[4.57, 9.56]	0.24	[0.13, 0.35]	1.27	1.42
Charlotte	50.9	350	1.81	[-0.89, 4.51]	0.01	[-0.14, 0.16]	1.01	1.21
Chicago	53.5	366	6.97	[2.88, 11.07]	0.29	[0.14, 0.44]	1.34	1.33
Dallas	38.7	4,233	5.38	[4.11, 6.64]	0.17	[0.12, 0.23]	1.19	1.31
Houston	40.6	2,179	5.59	[3.73, 7.45]	0.11	[0.04, 0.18]	1.12	1.34
Los Angeles	54.8	1,092	10.74	[7.14, 14.34]	0.31	[0.19, 0.43]	1.36	1.67
Milwaukee	38.9	1,343	2.97	[1.54, 4.39]	0.13	[0.06, 0.20]	1.14	1.17
New York	52.9	1,573	6.36	[3.51, 9.20]	0.19	[0.08, 0.29]	1.21	1.38
Rochester	45.0	344	2.73	[-0.18, 5.64]	0.23	[0.03, 0.42]	1.25	1.10
Sacramento	55.7	1,271	7.32	[5.08, 9.56]	0.21	[0.12, 0.29]	1.23	1.27
San Antonio	47.2	791	5.38	[3.12, 7.63]	0.15	[0.06, 0.25]	1.17	1.28
San Diego	63.4	942	6.34	[4.01, 8.66]	0.25	[0.17, 0.34]	1.29	1.33
San Francisco	57.8	598	9.89	[5.35, 14.44]	0.24	[0.12, 0.37]	1.28	1.50
San Jose	70.5	374	11.80	[7.25, 16.35]	0.49	[0.28, 0.69]	1.63	1.48

[1]Count-to-capacity ratio using weighted trip count and roadstock proxies in Section 5.3.
[2]Only for cities with n > 300.
[3]Average queueing delay by MSA estimated using model in Table 3, Column 5.
[4]Estimated using model in Table 3 Column 5 with log(commute time) as dependent variable.
[5]Exponential of Q (log). Ratio of average travel time with and without bottleneck.
[6]Average of quarterly travel time indices (TTI) sourced from FHWA. TTI is only available for MSAs.
[7]Estimates and confidence intervals calculated from Table 3, Columns (5) and (6).

Table 7: Average and total queueing cost in 2017.

	Pop (Millions)[1]	Commutes (Millions)[2]	AQC (2017$)[3]	TQC (Million$)[3]
USA	321.00	16,544.220	0.713	11,793.851
MSAs	179.00	9,480.905	1.221	11,580.648
Atlanta	5.70	328.150	1.719	564.078
Austin	2.00	149.594	1.549	231.763
Charlotte	2.43	136.607	0.397	54.266
Chicago	9.55	493.252	1.529	754.032
Dallas	7.10	410.812	1.178	484.100
Houston	6.64	323.581	1.226	396.775
Los Angeles	13.26	663.448	2.354	1,562.023
Milwaukee	1.58	97.765	0.650	63.563
New York	20.19	740.055	1.394	1,031.720
Rochester	1.08	57.869	0.599	34.655
Sacramento	2.27	131.989	1.604	211.720
San Antonio	2.38	140.336	1.179	165.444
San Diego	3.28	194.617	1.390	270.437
San Francisco	4.64	214.428	2.169	465.155
San Jose	1.97	113.220	2.587	292.917

[1]Sourced from Census data.
[2]Author's calculations using NHTS survey weights.
[3]Average queueing cost (AQC) and total queueing cost (TQC) calculated based on U.S. average hourly wages.

6 Social cost

6.1 Social cost of queueing delays

In this section, I quantify the welfare loss due to bottleneck queueing delays, and discuss how it can, in principle, be recovered by toll or congestion pricing. For 2017, I compute the average queueing cost (AQC) which is the per-commute dollar cost of the average bottleneck delay and then aggregate across all bottleneck-period trips to obtain the total queueing cost (TQC) borne by every morning bottleneck commuter.

To find AQC, an appropriate value of commute time, a, has to be chosen. Small and Verhoef (2007, pp. 52–53) find that empirical estimates of the value of time for personal travel span from 20 % to 90 % of the gross wage, with an average of roughly 50 %. Wolff (2014) empirically estimated a travel value of time that is 50% of the wage rate as well. Since the U.S Department of Transportation guidance in 2016 is also an adjustment of 50% from gross wage (U.S. Department of Transportation, 2016), I decide to adopt this adjustment factor. The 2017 U.S. annual hourly wage rate from the BLS CES is $26.31. Multiplying this with the adjustment factor, we get $a = \$0.22$ per minute.

Table 7 shows AQC derived by multiplying the $\bar{Q}$ estimates in Table 6 with a. TQC is found by multiplying AQC with the weighted sum number of trips per year from the commutes (millions) column. The 2017 social cost of queueing delays due to bottlenecks in the U.S. is around $11.8 billion, and the identified MSAs from the NHTS accounted for 98% of this cost. The top 3 MSAs by population in the table (New York, Los Angeles and Chicago) had the highest social cost. Atlanta is fourth on the TQC ranking but has a smaller population than Dallas and Houston. Both MSAs in the state of Texas had smaller N/ψ ratios and $\bar{Q}$ estimates than Atlanta in 2017.

The $11.8 billion[10] social cost estimate is small compared to the $179 billion national congestion cost reported by the aforementioned Urban Mobility Report (UMR, see Section 5.4). Again, it is important to differentiate what both values are measuring. The UMR reports both direct costs of whole-day congestion time and fuel wasted for all vehicles on the road measured through vehicle probe data, and indirect costs like increased logistical prices. In this study, only observations of morning commutes are used.

The AQC and TQC from this study is a measure of queue delay caused by bottleneck dynamics where commuter scheduling is endogenous. While TQC is the pure deadweight loss from bottleneck queues, the marginal social cost (MSC) at the bottleneck equilibrium needs to incorporate the cost from scheduling delays as well. In a no-toll equilibrium, ADL shows that MSC is four times that of the AQC. Recall that in Equation 16, the total queueing cost is equal to the total schedule cost. Summing both up gives us the total cost at equilibrium. The change in this total cost when one more additional commuter is added, or marginal social cost, is then shown to be twice of this total cost divided by N, giving ADL's result. To reiterate, TQC specifically measures welfare loss due to queueing but recovery of this welfare loss will need to consider scheduling.

6.2 Pricing bottlenecks

ADL describes a first-best time-varying toll within the bottleneck window that eliminates all queueing by ensuring arrivals at the bottleneck do not exceed capacity ψ. This is achieved by commuters redistributing their arrivals into a new toll equilibrium, where no

[10]The estimate of $11.8 billion depends on key assumptions, most notably the choice of value-of-time, commute observations, and count measures. A dedicated sensitivity analysis varying a and using more representatives trip observations would be straightforward for interested parties.

Table 8: Queueing delays and costs in New York by arrival time at work in 2017.

Arrived within	Commutes (%)	$\bar{Q}$ (mins)	95% CI	AQC (2017$)[1]	TQC (Millions$)[1]
6:00-11:00 AM	100.00	6.359	[3.51, 9.20]	1.862	1,377.904
6:00-6:30 AM	3.62	2.415	[-1.61, 6.44]	0.707	18.941
6:30-7:00 AM	6.24	4.118	[0.92, 7.31]	1.206	55.670
7:00-7:30 AM	9.43	4.497	[1.18, 7.82]	1.317	91.903
7:30-8:00 AM	18.61	5.204	[2.17, 8.24]	1.524	209.805
8:00-8:30 AM	17.47	6.690	[3.57, 9.81]	1.959	253.305
8:30-9:00 AM	17.54	7.096	[3.88, 10.32]	2.078	269.690
9:00-9:30 AM	13.62	9.052	[5.47, 12.63]	2.650	267.102
9:30-10:00 AM	5.97	9.360	[4.67, 14.05]	2.741	121.077
10:00-10:30 AM	4.86	7.601	[3.59, 11.61]	2.226	79.999
10:30-11:00 AM	2.65	9.112	[2.84, 15.38]	2.668	52.303

[1]Average queueing cost (AQC) and total queueing cost (TQC) calculated based on New York average hourly wages.

commuter has an incentive to alter their arrival pattern again as toll paid plus scheduling cost replicates the original queue-schedule cost. Since toll revenue equals TQC, welfare is recovered. With more recent advances like distance-based road pricing using satellite networks (Theseira, 2020), it may be technologically feasible to identify bottlenecks and charge a time-varying toll. However, a whole-of-U.S., MSA, or even route specific roll-out of such a system will face steep political economy hurdles, and privacy concerns, beyond mere technical feasibility (King et al., 2007; Whitty & Svadlenak, 2009).

ADL proposes a second-best coarse toll with a single fee within the bottleneck window. Queueing is reduced but not fully eliminated like in the time-varying toll, with the trade-off being practicality in implementation. Kim further suggests a middle ground where the toll prices can be set with respect to $\bar{Q}$ across arrival time bins (see Figure 4). Since a nation-wide introduction of such a toll is impractical, I illustrate this approach using an individual MSA. New York City introduced congestion pricing in 2025 (Metropolitan Transportation Authority, 2025), presenting an opportunity to compare the actual tolls set by NYC with AQC estimates. Also, unlike most MSAs, hourly wage for New York also publicly available for 2017 (Bureau of Labor Statistics, 2017).

Table 8 presents our estimates of $\bar{Q}$ for the New York MSA over the full 6–11 am bottleneck window and in 30-minute blocks. Column 2 reports each block's share of total trips. Employing a city-specific wage of $35.14 per hour ($a = 0.29$ per minute), we compute AQC and TQC exactly as in Section 6.1. The AQC for all commutes is $1.86, rising to $2.74 between 9:30 to 10:00am. That is $2.45 and $3.61 respectively, adjusted for June 2025 prices.

In 2025, the New York City congestion zone[11] toll is set at $9 for peak-period access (5:00am to 9:00pm), with downward adjustments depending on choice of entry and exit points. Splitting the charge equally on entry and exit implies a round-trip toll of $4.50 per direction which is still well above our highest AQC estimate. By this measure, it is reasonable to believe that the congestion price will have an impact on morning traffic. Yet, Cook et al. (2025) use Google Maps Traffic Trends with a synthetic-control design to show that congestion pricing regime only has a small, statistically insignificant morning speed effect, with larger gains in the afternoon and early evening. Again, this discrepancy may reflect methodology. The micro-level bottleneck estimates in this study exploit within-commuter variation to recover pure queueing delay for identical trip profiles, whereas Cook et. al. measure aggregate network speeds across all users. This suggests room to blend both approaches in future studies. Regardless, to improve the performance of congestion pricing in the morning, policymakers in New York can consider replacing the flat cordon fee with time-block varying prices calibrated to more recent $\bar{Q}$ profiles across major choke points. This can smooth arrivals where queues actually form without over-taxing periods of free-flow.

The recommendation needs to be framed by several caveats. A fully optimal toll hinges on knowing each commuter's, or each group of commuters', value-of-time associated parameters $[a, b, c]$, and their ideal start-work time t^*. In practice these parameters vary widely across individuals and cannot be directly observed, so any one-size-fits-all toll risks over- or under-charging many users. When tolls rise high enough, some commuters may simply drop out, shifting to off-peak travel, home office, or alternate modes. Thus, welfare analyses must account for participation elasticity and not just schedule-delay relief (Van Den Berg & Verhoef, 2011). Calibrating, coordinating and updating dozens of time-varying tolls based on $\bar{Q}$ profiles across a complex traffic network with many bottlenecks can also be very challenging. This and the aforementioned satellite-based road-usage tracking system may be feasible in a city-state like Singapore (Menon, 2023), but presents an institutional challenge for implementation in the U.S with its differing road authorities and legislative mandates (Federal Highway Administration, 2008a; Stockton, 2017).

If a deliberative pricing strategy hits these limits, it is still possible to try other levers to shrink peak-hour queues by reducing the count-to-capacity ratio, N/ψ beyond capital-intensive investments into infrastructure. Referring back to Section 5.2, employers can stag-

[11]The congestion zone is designated in Manhattan as south and inclusive of 60th Street.

ger start times or subsidize telecommuting to reduce the concentration of commutes into a bottleneck window. Using 2017-18 travel survey data from Austin, Texas, (Rahman et al., 2021) show that offering flextime reduces the probability of morning peak departures by 8.2 percentage points. Complementing long-term cross-modal improvements, transport authorities can also deploy ramp metering, reversible lanes or adaptive signal timing on key channels to temporarily raise throughput (Federal Highway Administration, 2014). Looking ahead, autonomous vehicles and mobility-as-a-service platforms may lower effective values of time and reshape scheduling incentive, but they also pose new coordination and equity challenges for any congestion policy (Fagnant & Kockelman, 2015; Filipe & Heath, 2023; Garikapati et al., 2017; Yu et al., 2022).

7 Discussion

This paper provides nationally representative, trip-level estimates of morning-peak queuing delays for 2001, 2009, and 2017 using the bottleneck framework, extending single-wave work to a three-wave design and mapping results to AQC/TQC at national and MSA levels. By recovering route-integrated queuing time from trip diaries rather than segment speeds, the estimates speak directly to the Vickrey/ADL bottleneck welfare loss approach, and complement corridor-based Pigouvian approaches that often require assumed demand curves and miss cumulative route delays.

The findings in this paper also align with recent evidence showing that observed delay profiles are flatter than the standard model predicts, consistent with the presence of infra-marginal commuters and heterogeneous scheduling preferences (Hall, 2024; Kim & Moon, 2024).

Relative to GPS/Google-Maps telemetry, our contribution is representativeness and structural interpretability. The nationally representative NHTS trip diaries recover the bottleneck's queuing component and link it to MSA-level determinants with rich household covariates. Telemetry studies, by contrast, trade breadth for temporal resolution.

Used together, these strands are complementary for policy. Telemetry validates free-flow baselines, sharpens intra-peak gradients for block-pricing design, and measures within-day heterogeneity, while NHTS supports welfare accounting and distributional inference. Since unit-level linkage of telemetry to survey microdata raises re-identification risk, I work only

with publicly available data with privacy-preserving aggregates. Future studies can pursue secure-enclave or trusted-research-environment arrangements to enable such linkage while safeguarding confidentiality.

In the canonical model, average queuing time rises proportionally with the count-to-capacity ratio N/ψ and with the value-of-time composite $bc/[a(b+c)]$. My determinants analysis line up with these comparative statics. Higher trip volumes and wages associate with larger delays. VMT outperforms raw commute counts as a demand proxy, and exclusive-lane expansions show mixed signs consistent with capacity reallocation rather than net capacity growth.

ADL show that a time-varying toll within the bottleneck window decentralizes the social optimum by eliminating queues. Toll revenue equals TQC, so welfare is fully recovered. This is the policy ideal that practical instruments approximate. Administrative and political constraints push cities toward coarser instruments. ADL show that step tolls capture a substantial share of first-best gains. Two second-best options map directly from my arrival-profile results. First, a block (step) pricing by 15–30-minute windows calibrated to observed delay gradients. Second, cordon pricing with peak premiums captures multiple bottlenecks. Taken together, charging vehicles to enter or drive within a defined zone during set hours can be implemented as time-block prices aligned to observed delay gradients.

Where pricing is partial or delayed, staggered work hours, telework incentives, ramp metering, reversible lanes, and adaptive signals can help reduce N/ψ. Staggered-hours policy merits distinct attention. In bottleneck models with production externalities (firms choose work-start times), Pigouvian first-best is locally stable, whereas second-best regimes that cannot eliminate queues can be unstable and exhibit multiple equilibria, cautioning against coarse pricing without coordinated start-time policy (Takayama, 2015).

The estimates describe pre-COVID commuting equilibria and serve as a clean baseline for assessing persistent post-COVID shifts in desired-arrival distributions and telework. In 2020, annual delay per auto commuter fell from 54 hours (2019) to 27 hours (–50%), and both the daily and weekly congestion profiles flattened, as documented via a traffic telemetry approach (Schrank et al., 2024). By 2022, congestion rebounded unevenly. Nationally, levels were generally below 2019 but rising, with patterns ‘more spread-out’ across the day and week; the evening peak was returning, while the morning peak remained shorter and smaller than before COVID.

The bottleneck logic is unchanged. Queues form when demand (or arrival at the bottleneck), N, exceeds capacity ψ. COVID and Post-COVID shifts, via work-from-home or sectoral changes (Dingel & Neiman, 2020), can reduce N. This leads to a reduction in both the peak queue $q_{\max}$ and average queuing time $\bar{Q} \propto \frac{bc}{a(b+c)} \cdot \frac{N}{\psi}$. A spreading of arrival-at-work times directly impacts arrival rates at the bottleneck so fewer commuters queue and total queuing cost (TQC) falls. Shifts in a, b, c may occur through wages or schedule norms but, absent large wage movements, they are second-order relative to changes in the demand-capacity ratio. The policy mapping therefore stays the same. Time-varying tolls and staggered hours target demand where it exceeds capacity. Only the calibration—block edges and levels need to adjust to the new arrival profile.

The estimates in this study are a lower bound on welfare losses. With heterogeneous scheduling costs, the first-best can improve welfare not only by eliminating queues, but also by rearranging arrival times to better suit individual preferences, thereby reducing schedule delays. This is a a standard implication of the bottleneck model (Arnott et al., 1994). Since I isolate the queuing wedge, I do not account for other externalities (e.g., fuel use, local air pollution, carbon emissions, crash risk, noise, reliability). Time-varying tolls that eliminate queues nonetheless mitigate several of these harms by shortening time in traffic and smoothing speeds. Pairing congestion pricing with emissions or fuel charges can address residual pollution externalities (Small & Verhoef, 2007, p. 171).

Sensitivity to arrival-window cutoffs is mechanical in the design of the empirical framework in this study. Cutoff selection by difference-in-means, queueing time profiles and within-household checks mitigate arbitrariness. Free-flow times and effective capacity are not directly observed and are proxied rather than measured. The trip-pair design addresses this by comparing identical routes across arrival windows. Mixed HOV/HOT associations likely reflect lane reallocation without net increases in bottleneck capacity. Future DiD/Synth designs around lane additions, with consideration for MSA-level traffic pattern changes, can separate reallocation from expansion effects.

Anchored in the bottleneck prinicipal that peak queues are pure deadweight loss, our estimates provide sufficient statistics for action. First, set time-block tolls that mirror observed delay gradients. Second, pair pricing with staggered work hours and signal timing/ramp-metering where pricing is partial, and evaluate with synthetic controls as deployments scale. Together these instruments target exactly the welfare wedge our

AQC/TQC quantify, with calibration (block edges and levels) updated as the arrival profile shifts.

8 Conclusion

This paper quantifies the deadweight loss from morning commuter bottlenecks in the United States using nationally representative trip-level data and a structural bottleneck framework. I show that the U.S.morning-peak bottlenecks impose persistent welfare losses across 2001, 2009, and 2017, rising to $11.8 billion annually by the latest wave. These losses scale with metropolitan demand–capacity ratios. Cross-MSA patterns reveal that delays grow with vehicle miles travelled and per-capita income, and decline where toll-lane capacity has expanded, consistent with the model's comparative statics.

By recovering route-integrated queueing time from trip diaries, the analysis complements high-resolution telemetry studies while offering structural interpretability and distributional insight. I argue that the estimates of average queueing profile for a city provide a core metrics for designing second-best interventions. Time-block tolls calibrated to observed delay gradients, paired with staggered work hours and network management tools, can target the welfare wedge even when first-best continuous tolls are impractical.

The estimated welfare losses are a lower bound, as heterogeneous scheduling preferences open scope for further gains through re-ordering arrivals. Beyond congestion pricing, complementary policies such as emissions charges, telework incentives, or start-time coordination can address residual externalities and reinforce efficiency gains.

Looking forward, the same bottleneck logic applies as travel behaviour evolves with new technologies. The deployment of autonomous vehicles, examined will interact with congestion both directly, through changes in flow and capacity, and indirectly, through commuters' willingness to delegate driving. Understanding these interactions is essential for anticipating whether AV adoption will amplify or alleviate the queueing costs quantified here.

Appendix

The appendix contains estimations of average queueing time from 2001, 2009, 2017 and 2022 using an alternative specification with Census Division controls and fixed effects. The determinants estimation is conducted using the census-derived number of commuters from MSAs instead of trips from the NHTS. Finally, I provide the city-level measures for 2001 and 2009.

Table 9: Estimates of average queueing time within bottlenecks.

	Commute time				
	(1)	(2)	(3)	(4)	(5)
Bottleneck window	2.85***	1.74***	1.56***	1.57***	1.29***
	(0.097)	(0.189)	(0.189)	(0.190)	(0.187)
Bottleneck window × 2009		0.766***	0.763***	0.863***	1.06***
		(0.239)	(0.241)	(0.240)	(0.240)
Bottleneck window × 2017		1.82***	1.81***	1.82***	1.94***
		(0.252)	(0.253)	(0.253)	(0.251)
Bottleneck window × 2022		2.31**	2.16**	2.08*	2.14**
		(1.06)	(1.05)	(1.07)	(1.07)
Distance variables	✓	✓	✓	✓	✓
Area variables			✓	✓	✓
Household/commuter variables				✓	✓
Observations	114,430	114,430	114,430	114,430	114,430
R^2	0.727	0.732	0.738	0.740	0.747
Day-of-week fixed effects					✓
Division fixed effects					✓

OLS regressions with a constant from (1) to (4). OLS fixed effects for (5). Heteroskedasticity-robust standard errors in parentheses. $***p < 0.01, **p < 0.05, *p < 0.1$.

Table 10: Estimates of proxies (commuters) and covariates of bottleneck congestion.

	Commute time			
	(1)	(2)	(3)	(4)
Panel A	Proxies	log(AWW)	log(AAW)	log(TAW)
log(Commuters)	2.25**	0.160	0.160	
	(0.911)	(0.996)	(0.996)	
log(FWY Lane Miles)	-0.656	-0.102	-0.102	-0.339
	(0.526)	(0.529)	(0.529)	(0.493)
log(MRU Lane Miles)	-0.409	0.469	0.469	0.207
	(0.824)	(0.834)	(0.834)	(0.508)
Added variable		6.17***	6.17***	1.46***
		(1.47)	(1.47)	(0.356)
Observations	49,150	49,150	49,150	49,150
R^2	0.705	0.705	0.705	0.704

	Commute time			
	(1)	(2)	(3)	(4)
Panel B	Sprawl	Centeredness	Use Mix	Density
log(Commuters)	0.091	0.737	-0.013	-0.073
	(1.23)	(1.06)	(1.14)	(1.18)
log(FWY Lane Miles)	0.607	0.342	0.496	0.422
	(0.685)	(0.625)	(0.635)	(0.622)
log(MRU Lane Miles)	0.830	0.618	0.742	0.733
	(1.05)	(1.02)	(1.02)	(1.02)
Added variable	0.017	0.018	0.029**	0.025**
	(0.012)	(0.014)	(0.012)	(0.012)
Observations	27,700	27,700	27,700	27,700
R^2	0.731	0.731	0.731	0.731

	Commute time			
	(1)	(2)	(3)	(4)
Panel C	Network	log(VMT)	log(Toll)	log(HOV)
log(Commuters)	0.552	-2.69**	2.32**	2.15**
	(1.10)	(1.26)	(0.907)	(0.912)
log(FWY Lane Miles)	0.678	-2.12***	-1.10**	-0.571
	(0.654)	(0.581)	(0.547)	(0.531)
log(MRU Lane Miles)	0.063	-2.46***	0.238	-0.619
	(1.00)	(0.899)	(0.851)	(0.838)
Added variable	0.020*	7.86***	-0.246***	0.169
	(0.011)	(1.31)	(0.093)	(0.107)

Table 11: Bottleneck congestion by cities in 2001.

	N/ψ[1]	n[2]	$\bar{Q}$ (mins)[3]	95% CI	Q (log)[4]	95% CI	Q-Index[5]
USA[6]	41.2	112,745	1.45	[1.09, 1.82]	0.05	[0.03, 0.07]	1.05
MSAs[6]	49.8	49,150	3.14	[2.22, 4.06]	0.10	[0.06, 0.14]	1.10
Boston	146.5	249	6.10	[3.39, 8.82]	0.47	[0.23, 0.71]	1.60
Buffalo	43.4	237	1.69	[-0.60, 3.97]	0.13	[-0.10, 0.35]	1.13
Chicago	58.0	327	4.04	[0.68, 7.40]	0.20	[0.07, 0.33]	1.22
Los Angeles	49.5	428	2.12	[-2.03, 6.26]	0.09	[-0.10, 0.29]	1.10
Milwaukee	51.4	547	1.13	[-0.24, 2.50]	0.01	[-0.08, 0.10]	1.01
Minneapolis	73.9	221	0.20	[-3.52, 3.92]	0.01	[-0.15, 0.17]	1.01
New York	61.1	1,819	4.82	[1.90, 7.75]	0.17	[0.05, 0.29]	1.18
Rochester	63.4	439	1.25	[-1.33, 3.83]	0.00	[-0.13, 0.12]	1.00
San Francisco	58.6	224	5.67	[1.63, 9.72]	0.22	[0.06, 0.39]	1.25
Washington	48.6	270	1.58	[-3.61, 6.77]	0.04	[-0.12, 0.20]	1.04

[1]Count-to-capacity ratio using weighted trip count and roadstock proxies in Section 5.3.

[2]Only for cities with n $>$ 200.

[3]Average queueing delay by MSA estimated using model in Table 3, Column 5.

[4]Estimated using model in Table 3 Column 5 with log(commute time) as dependent variable.

[5]Exponential of Q (log). Ratio of average travel time with and without bottleneck.

[6]Estimates and confidence intervals calculated from Table 3, Columns (5) and (6).

Table 12: Bottleneck congestion by cities in 2009.

	N/ψ[1]	n[2]	$\bar{Q}$ (mins)[3]	95% CI	Q (log)[4]	95% CI	Q-Index[5]
USA[6]	41.2	112,745	3.25	[2.92, 3.58]	0.11	[0.10, 0.13]	1.12
MSAs[6]	49.8	49,150	5.57	[5.03, 6.11]	0.18	[0.16, 0.20]	1.20
Atlanta	60.6	423	4.24	[-0.57, 9.04]	0.11	[-0.05, 0.28]	1.12
Austin	49.0	538	1.60	[-4.55, 7.75]	0.11	[-0.14, 0.35]	1.11
Dallas	38.0	2,082	4.84	[3.38, 6.29]	0.17	[0.11, 0.23]	1.19
Houston	38.3	1,393	4.78	[3.08, 6.48]	0.18	[0.09, 0.27]	1.20
Jacksonville	45.2	362	4.44	[1.45, 7.43]	0.17	[0.05, 0.28]	1.18
Los Angeles	50.0	1,466	6.15	[4.38, 7.93]	0.19	[0.12, 0.25]	1.20
Miami	60.4	1,069	5.38	[2.73, 8.03]	0.20	[0.08, 0.32]	1.22
New York	52.1	1,245	5.26	[2.70, 7.83]	0.21	[0.11, 0.31]	1.24
Norfolk	45.1	1,120	1.20	[-1.34, 3.74]	0.02	[-0.07, 0.12]	1.02
Orlando	47.9	358	3.21	[-0.64, 7.06]	0.08	[-0.07, 0.23]	1.08
Richmond	38.3	817	1.60	[-1.57, 4.77]	0.09	[-0.05, 0.23]	1.09
Riverside	29.7	464	5.48	[1.84, 9.12]	0.17	[0.06, 0.28]	1.19
Sacramento	55.2	373	1.34	[-0.95, 3.63]	0.02	[-0.08, 0.13]	1.02
San Antonio	57.2	681	1.93	[-0.20, 4.07]	0.04	[-0.06, 0.14]	1.04
San Diego	49.9	1,839	3.01	[1.62, 4.39]	0.09	[0.03, 0.15]	1.09
San Francisco	58.3	635	7.51	[4.03, 11.00]	0.24	[0.11, 0.38]	1.27
Tampa	54.2	533	2.73	[-0.56, 6.02]	0.02	[-0.14, 0.18]	1.02
Washington	64.1	781	6.75	[3.55, 9.95]	0.14	[0.06, 0.22]	1.15

[1]Count-to-capacity ratio using weighted trip count and roadstock proxies in Section 5.3.

[2]Only for cities with n > 300.

[3]Average queueing delay by MSA estimated using model in Table 3, Column 5.

[4]Estimated using model in Table 3 Column 5 with log(commute time) as dependent variable.

[5]Exponential of Q (log). Ratio of average travel time with and without bottleneck.

[6]Estimates and confidence intervals calculated from Table 3, Columns (5) and (6).

References

Akbar, P., Couture, V., Duranton, G., & Storeygard, A. (2023). Mobility and Congestion in Urban India. *American Economic Review*, *113*(4), 1083–1111. https://doi.org/10.1257/aer.20181662

Arnott, R., De Palma, A., & Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics*, *27*(1), 111–130. https://doi.org/10.1016/0094-1190(90)90028-L

Arnott, R., de Palma, A., & Lindsey, R. (1993). A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. *American Economic Review*, *83*(1), 161–179.

Arnott, R., de Palma, A., & Lindsey, R. (1994). The Welfare Effects of Congestion Tolls with Heterogeneous Commuters. *Journal of Transport Economics and Policy*, *28*(2), 139–161.

Brent, D. A., & Gross, A. (2018). Dynamic road pricing and the value of time and reliability. *Journal of Regional Science*, *58*(2), 330–349. https://doi.org/10.1111/jors.12362

Bricka, S., Reuscher, T., Schroeder, P., Fisher, M., Beard, J., & Sun, L. (2024). *Summary of travel trends: 2022 national household travel survey* (Technical Report). U.S. Department of Transportation, Federal Highway Administration. https://nhts.ornl.gov/media/2022/pub/2022%5C_NHTS%5C_Summary%5C_Travel%5C_Trends.pdf

Brownstone, D., & Small, K. A. (2005). Valuing time and reliability: Assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy and Practice*, *39*(4), 279–293. https://doi.org/10.1016/j.tra.2004.11.001

Bureau of Labor Statistics. (2017). Current employment statistics – national (ces-n) [Data retrieved August 1, 2025].

Bureau of Labor Statistics. (2025, May). Quarterly census of employment and wages (qcew) [Accessed 2025-05-06]. https://doi.org/10.3886/E226867V1

Cohen, M. C., Jacquillat, A., Ratzon, A., & Sasson, R. (2022). The impact of high-occupancy vehicle lanes on carpooling. *Transportation Research Part A: Policy and Practice*, *165*, 186–206. https://doi.org/10.1016/j.tra.2022.08.021

Cook, C., Kreidieh, A., Vasserman, S., Allcott, H., Arora, N., van Sambeek, F., Tomkins, A., & Turkel, E. (2025, March). *The short-run effects of congestion pricing in new york city* (Working Paper No. 33584). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w33584

Couture, V., Duranton, G., & Turner, M. A. (2018). Speed. *The Review of Economics and Statistics*, *100*(4), 725–739. https://doi.org/10.1162/rest_a_00744

Daganzo, C. F., & Cassidy, M. J. (2008). Effects of high occupancy vehicle lanes on freeway congestion. *Transportation Research Part B: Methodological*, *42*(10), 861–872. https://doi.org/10.1016/j.trb.2008.03.002

Dewees, D. N. (1979). Estimating the Time Costs of Highway Congestion. *Econometrica*, *47*(6), 1499–1512. https://doi.org/10.2307/1914014

Dingel, J. I., & Neiman, B. (2020). How many jobs can be done at home? *Journal of Public Economics*, *189*, 104235. https://doi.org/10.1016/j.jpubeco.2020.104235

Dumouchel, W. H., & Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, *78*(383), 535–543. https://doi.org/10.1080/01621459.1983.10478006

Duranton, G., & Turner, M. A. (2011). The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review*, *101*(6), 2616–2652. https://doi.org/10.1257/aer.101.6.2616

Ewing, R., & Cervero, R. (2010). Travel and the Built Environment. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Ewing, R., Pendall, R., & Chen, D. (2003). Measuring Sprawl and Its Transportation Impacts. *Transportation Research Record*, *1831*(1), 175–183. https://doi.org/10.3141/1831-20

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Federal Highway Administration. (2008a). *Congestion Pricing: A Primer* (tech. rep. No. FHWA-HOP-08-039). U.S. Department of Transportation. https://ops.fhwa.dot.gov/publications/fhwahop08039/fhwahop08039.pdf

Federal Highway Administration. (2008b). *Hpms reassessment 2010+ final report* (tech. rep.). U.S. Department of Transportation, Federal Highway Administration. Washington, DC. https://www.fhwa.dot.gov/policyinformation/hpms/2010/hpms2010.pdf

Federal Highway Administration. (2009). *A compendium of existing hov lane facilities in the united states* (tech. rep. No. FHWA-HOP-03-030) (Accessed 2025-05-06). U.S.

Department of Transportation. Washington, DC. https://rosap.ntl.bts.gov/view/dot/865

Federal Highway Administration. (2014). *Transportation Systems Management and Operations Benefit–Cost Analysis Compendium* (tech. rep. No. FHWA-HOP-14-032). U.S. Department of Transportation. https://ops.fhwa.dot.gov/publications/fhwahop14032/fhwahop14032.pdf

Federal Highway Administration. (2016). *Highway performance monitoring system (hpms) field manual* (Field Manual No. 2125-0028). U.S. Department of Transportation. Washington, DC.

Federal Highway Administration. (2018). *2017 Urban Congestion Trends* (tech. rep. No. FHWA-HOP-18-025). U.S. Department of Transportation. Washington, DC. https://ops.fhwa.dot.gov/publications/fhwahop18025

Federal Highway Administration. (2024). *2024 national inventory of specialty lanes and roads* (tech. rep. No. FHWA-HOP-24-006) (Accessed 2025-05-06). U.S. Department of Transportation. Washington, DC. https://ops.fhwa.dot.gov/publications/fhwahop24006/fhwahop24006.pdf

Filipe, R. P., & Heath, A. (2023). Impacts of implementing mobility as a service in urban areas – a systematic literature review. *Transportation Research Procedia*, *72*, 179–186. https://doi.org/10.1016/j.trpro.2023.11.392

Fosgerau, M., & Small, K. A. (2012). Marginal Congestion Cost on a Dynamic Expressway Network. *Journal of Transport Economics and Policy*, *46*(3), 431–450.

Garikapati, V. M., Dias, F. F., Lavieri, P. S., Astroza, S., Pendyala, R. M., & Bhat, C. R. (2017). A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation*, *44*(6), 1307–1323. https://doi.org/10.1007/s11116-017-9797-8

Geroliminis, N., & Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, *42*(9), 759–770. https://doi.org/10.1016/j.trb.2008.02.002

Glaeser, E. L., Kahn, M. E., Arnott, R., & Mayer, C. (2001). Decentralized Employment and the Transformation of the American City [with Comments]. *Brookings-Wharton Papers on Urban Affairs*, 1–63.

Goel, R., & Burris, M. W. (2012). Hot lane policies and their implications. *Transportation*, *39*(6), 1019–1033. https://doi.org/10.1007/s11116-011-9382-5

Hall, J. D. (2024). Inframarginal Travelers and Transportation Policy. *International Economic Review*, *65*(3), 1519–1550. https://doi.org/10.1111/iere.12692

Hamidi, S., Ewing, R., Preuss, I., & Dodds, A. (2015). Measuring Sprawl and Its Impacts: An Update. *Journal of Planning Education and Research*, *35*(1), 35–50. https://doi.org/10.1177/0739456X14565247

Kim, J. (2019). Estimating the social cost of congestion using the bottleneck model. *Economics of Transportation*, *19*, 100119. https://doi.org/10.1016/j.ecotra.2019.100119

Kim, J., & Moon, J. (2024). Congestion and scheduling preferences of car commuters in California: Estimates using big data. *Journal of Economic Geography*, *24*(1), 145–170. https://doi.org/10.1093/jeg/lbad033

King, D., Manville, M., & Shoup, D. (2007). The political calculus of congestion pricing. *Transport Policy*, *14*(2), 111–123. https://doi.org/10.1016/j.tranpol.2006.11.002

Kreindler, G. (2024). Peak-Hour Road Congestion Pricing: Experimental Evidence and Equilibrium Implications. *Econometrica*, *92*(4), 1233–1268. https://doi.org/10.3982/ECTA18422

Li, Z.-C., Huang, H.-J., & Yang, H. (2020). Fifty years of the bottleneck model: A bibliometric review and future research directions. *Transportation Research Part B: Methodological*, *139*, 311–342. https://doi.org/10.1016/j.trb.2020.06.009

Manson, S., Schroeder, J., Van Riper, D., Knowles, K., Kugler, T., Roberts, F., & Ruggles, S. (2024). IPUMS National Historical Geographic Information System: Version 19.0. https://doi.org/10.18128/D050.V19.0

Menon, A. P. G. (2023). Commentary: Erp 2.0 — what is it good for and why do we need it? [Updated January 18, 2025]. *Channel NewsAsia*. https://www.channelnewsasia.com/today/voices/commentary-erp-20-what-it-good-and-why-do-we-need-it-4865821

Metropolitan Transportation Authority. (2025). Central business district tolling program [Program effective January 5, 2025]. https://www.mta.info/project/CBDTP

Needell, E., Amirgholy, V., & Asudegi, M. (2023). Estimating macroscopic fundamental diagrams for u.s. urban networks using probe-vehicle data and machine learning. *Transportation Research Part B: Methodological*, *158*, 99–115. https://doi.org/10.1016/j.trb.2023.08.003

Pigou, A. C. (1920). *The economics of welfare*. Macmillan; Co.

Rahman, M., Gurumurthy, K. M., & Kockelman, K. M. (2021). Impact of flextime on departure time choice for home-based commuting trips in austin, texas. *Transportation Research Record: Journal of the Transportation Research Board*, *2676*(1), 446–459. https://doi.org/10.1177/03611981211035757

Schrank, D., Albert, L., Jha, K., & Eisele, B. (2024, June). *2023 Urban Mobility Report* (tech. rep.). Texas A&M Transportation Institute, The Texas A&M University System. https://mobility.tamu.edu/umr/report/

Schrank, D., Eisele, B., & Lomax, T. (2019, August). *2019 urban mobility report* (tech. rep.) (Sponsored by the Texas Department of Transportation). Texas A&M Transportation Institute, Texas A&M University System. College Station, Texas. https://mobility.tamu.edu/umr/

Small, K. A. (1982). The Scheduling of Consumer Activities: Work Trips. *The American Economic Review*, *72*(3), 467–479.

Small, K. A. (2015). The bottleneck model: An assessment and interpretation. *Economics of Transportation*, *4*(1), 110–117. https://doi.org/10.1016/j.ecotra.2015.01.001

Small, K. A., & Verhoef, E. T. (2007). *The economics of urban transportation*. Routledge. https://doi.org/10.4324/9781315157375

Stockton, N. (2017). Why can't america have a single tolling system? because bureaucracy. *WIRED*. https://www.wired.com/story/national-tolling-system/

Takayama, Y. (2015). Bottleneck Congestion and Distribution of Work Start Times: The Economics of Staggered Work Hours Revisited. *Transportation Research Procedia*, *7*, 499–518. https://doi.org/10.1016/j.trpro.2015.06.026

Theseira, W. (2020). *Congestion control in singapore* (Discussion Paper No. 2020/10). International Transport Forum. Paris, OECD Publishing. https://doi.org/10.1787/7d266609-en

U.S. Department of Transportation. (2016, September). *Revised departmental guidance on valuation of travel time in economic analysis* (tech. rep.) (Issued September 27; Effective December 16, 2016). U.S. Department of Transportation, Office of the Secretary. Washington, DC. https://www.transportation.gov/sites/dot.gov/files/docs/2016%5C%20Revised%5C%20Value%5C%20of%5C%20Travel%5C%20Time%5C%20Guidance.pdf

Van Den Berg, V., & Verhoef, E. T. (2011). Winning or losing from dynamic bottleneck congestion pricing? *Journal of Public Economics*, *95*(7-8), 983–992. https://doi.org/10.1016/j.jpubeco.2010.12.003

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. https://doi.org/10.2307/1911814

Whitty, J. M., & Svadlenak, J. R. (2009). *Discerning the pathway to implementation of a national mileage-based charging system* (Special Report No. 299). Transportation Research Board of the National Academies. Washington, D.C. https://onlinepubs.trb.org/onlinepubs/sr/SR299Mileage.pdf

Wolff, H. (2014). Value of time: Speeding behavior and gasoline prices. *Journal of Environmental Economics and Management*, *67*(1), 71–88. https://doi.org/10.1016/j.jeem.2013.11.002

Yang, J., Purevjav, A.-O., & Li, S. (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. https://doi.org/10.1257/pol.20170195

Yu, X., Van Den Berg, V. A., & Verhoef, E. T. (2022). Autonomous cars and activity-based bottleneck model: How do in-vehicle activities determine aggregate travel patterns? *Transportation Research Part C: Emerging Technologies*, *139*, 103641. https://doi.org/10.1016/j.trc.2022.103641

Zhu, P. (2013). Telecommuting, household commute and location choice. *Urban Studies*, *50*(12), 2441–2459. https://doi.org/10.1177/0042098012474520

References

Akbar, P., Couture, V., Duranton, G., & Storeygard, A. (2023). Mobility and Congestion in Urban India. *American Economic Review*, *113*(4), 1083–1111. https://doi.org/10.1257/aer.20181662

Arnott, R., De Palma, A., & Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics*, *27*(1), 111–130. https://doi.org/10.1016/0094-1190(90)90028-L

Arnott, R., de Palma, A., & Lindsey, R. (1993). A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. *American Economic Review*, *83*(1), 161–179.

Arnott, R., de Palma, A., & Lindsey, R. (1994). The Welfare Effects of Congestion Tolls with Heterogeneous Commuters. *Journal of Transport Economics and Policy*, *28*(2), 139–161.

Brent, D. A., & Gross, A. (2018). Dynamic road pricing and the value of time and reliability. *Journal of Regional Science*, *58*(2), 330–349. https://doi.org/10.1111/jors.12362

Bricka, S., Reuscher, T., Schroeder, P., Fisher, M., Beard, J., & Sun, L. (2024). *Summary of travel trends: 2022 national household travel survey* (Technical Report). U.S. Department of Transportation, Federal Highway Administration. https://nhts.ornl.gov/media/2022/pub/2022%5C_NHTS%5C_Summary%5C_Travel%5C_Trends.pdf

Brownstone, D., & Small, K. A. (2005). Valuing time and reliability: Assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy and Practice*, *39*(4), 279–293. https://doi.org/10.1016/j.tra.2004.11.001

Bureau of Labor Statistics. (2017). Current employment statistics – national (ces-n) [Data retrieved August 1, 2025].

Bureau of Labor Statistics. (2025, May). Quarterly census of employment and wages (qcew) [Accessed 2025-05-06]. https://doi.org/10.3886/E226867V1

Cohen, M. C., Jacquillat, A., Ratzon, A., & Sasson, R. (2022). The impact of high-occupancy vehicle lanes on carpooling. *Transportation Research Part A: Policy and Practice*, *165*, 186–206. https://doi.org/10.1016/j.tra.2022.08.021

Cook, C., Kreidieh, A., Vasserman, S., Allcott, H., Arora, N., van Sambeek, F., Tomkins, A., & Turkel, E. (2025, March). *The short-run effects of congestion pricing in new york city* (Working Paper No. 33584). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w33584

Couture, V., Duranton, G., & Turner, M. A. (2018). Speed. *The Review of Economics and Statistics*, *100*(4), 725–739. https://doi.org/10.1162/rest_a_00744

Daganzo, C. F., & Cassidy, M. J. (2008). Effects of high occupancy vehicle lanes on freeway congestion. *Transportation Research Part B: Methodological*, *42*(10), 861–872. https://doi.org/10.1016/j.trb.2008.03.002

Dewees, D. N. (1979). Estimating the Time Costs of Highway Congestion. *Econometrica*, *47*(6), 1499–1512. https://doi.org/10.2307/1914014

Dingel, J. I., & Neiman, B. (2020). How many jobs can be done at home? *Journal of Public Economics*, *189*, 104235. https://doi.org/10.1016/j.jpubeco.2020.104235

Dumouchel, W. H., & Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, *78*(383), 535–543. https://doi.org/10.1080/01621459.1983.10478006

Duranton, G., & Turner, M. A. (2011). The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review*, *101*(6), 2616–2652. https://doi.org/10.1257/aer.101.6.2616

Ewing, R., & Cervero, R. (2010). Travel and the Built Environment. *Journal of the American Planning Association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Ewing, R., Pendall, R., & Chen, D. (2003). Measuring Sprawl and Its Transportation Impacts. *Transportation Research Record*, *1831*(1), 175–183. https://doi.org/10.3141/1831-20

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Federal Highway Administration. (2008a). *Congestion Pricing: A Primer* (tech. rep. No. FHWA-HOP-08-039). U.S. Department of Transportation. https://ops.fhwa.dot.gov/publications/fhwahop08039/fhwahop08039.pdf

Federal Highway Administration. (2008b). *Hpms reassessment 2010+ final report* (tech. rep.). U.S. Department of Transportation, Federal Highway Administration. Washington, DC. https://www.fhwa.dot.gov/policyinformation/hpms/2010/hpms2010.pdf

Federal Highway Administration. (2009). *A compendium of existing hov lane facilities in the united states* (tech. rep. No. FHWA-HOP-03-030) (Accessed 2025-05-06). U.S. Department of Transportation. Washington, DC. https://rosap.ntl.bts.gov/view/dot/865

Federal Highway Administration. (2014). *Transportation Systems Management and Operations Benefit–Cost Analysis Compendium* (tech. rep. No. FHWA-HOP-14-032). U.S. Department of Transportation. https://ops.fhwa.dot.gov/publications/fhwahop14032/fhwahop14032.pdf

Federal Highway Administration. (2016). *Highway performance monitoring system (hpms) field manual* (Field Manual No. 2125-0028). U.S. Department of Transportation. Washington, DC.

Federal Highway Administration. (2018). *2017 Urban Congestion Trends* (tech. rep. No. FHWA-HOP-18-025). U.S. Department of Transportation. Washington, DC. https://ops.fhwa.dot.gov/publications/fhwahop18025

Federal Highway Administration. (2024). *2024 national inventory of specialty lanes and roads* (tech. rep. No. FHWA-HOP-24-006) (Accessed 2025-05-06). U.S. Department of Transportation. Washington, DC. https://ops.fhwa.dot.gov/publications/fhwahop24006/fhwahop24006.pdf

Filipe, R. P., & Heath, A. (2023). Impacts of implementing mobility as a service in urban areas – a systematic literature review. *Transportation Research Procedia*, *72*, 179–186. https://doi.org/10.1016/j.trpro.2023.11.392

Fosgerau, M., & Small, K. A. (2012). Marginal Congestion Cost on a Dynamic Expressway Network. *Journal of Transport Economics and Policy*, *46*(3), 431–450.

Garikapati, V. M., Dias, F. F., Lavieri, P. S., Astroza, S., Pendyala, R. M., & Bhat, C. R. (2017). A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation*, *44*(6), 1307–1323. https://doi.org/10.1007/s11116-017-9797-8

Geroliminis, N., & Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, *42*(9), 759–770. https://doi.org/10.1016/j.trb.2008.02.002

Glaeser, E. L., Kahn, M. E., Arnott, R., & Mayer, C. (2001). Decentralized Employment and the Transformation of the American City [with Comments]. *Brookings-Wharton Papers on Urban Affairs*, 1–63.

Goel, R., & Burris, M. W. (2012). Hot lane policies and their implications. *Transportation*, *39*(6), 1019–1033. https://doi.org/10.1007/s11116-011-9382-5

Hall, J. D. (2024). Inframarginal Travelers and Transportation Policy. *International Economic Review*, *65*(3), 1519–1550. https://doi.org/10.1111/iere.12692

Hamidi, S., Ewing, R., Preuss, I., & Dodds, A. (2015). Measuring Sprawl and Its Impacts: An Update. *Journal of Planning Education and Research*, *35*(1), 35–50. https://doi.org/10.1177/0739456X14565247

Kim, J. (2019). Estimating the social cost of congestion using the bottleneck model. *Economics of Transportation*, *19*, 100119. https://doi.org/10.1016/j.ecotra.2019.100119

Kim, J., & Moon, J. (2024). Congestion and scheduling preferences of car commuters in California: Estimates using big data. *Journal of Economic Geography*, *24*(1), 145–170. https://doi.org/10.1093/jeg/lbad033

King, D., Manville, M., & Shoup, D. (2007). The political calculus of congestion pricing. *Transport Policy*, *14*(2), 111–123. https://doi.org/10.1016/j.tranpol.2006.11.002

Kreindler, G. (2024). Peak-Hour Road Congestion Pricing: Experimental Evidence and Equilibrium Implications. *Econometrica*, *92*(4), 1233–1268. https://doi.org/10.3982/ECTA18422

Li, Z.-C., Huang, H.-J., & Yang, H. (2020). Fifty years of the bottleneck model: A bibliometric review and future research directions. *Transportation Research Part B: Methodological*, *139*, 311–342. https://doi.org/10.1016/j.trb.2020.06.009

Manson, S., Schroeder, J., Van Riper, D., Knowles, K., Kugler, T., Roberts, F., & Ruggles, S. (2024). IPUMS National Historical Geographic Information System: Version 19.0. https://doi.org/10.18128/D050.V19.0

Menon, A. P. G. (2023). Commentary: Erp 2.0 — what is it good for and why do we need it? [Updated January 18, 2025]. *Channel NewsAsia*. https://www.channelnewsasia.com/today/voices/commentary-erp-20-what-it-good-and-why-do-we-need-it-4865821

Metropolitan Transportation Authority. (2025). Central business district tolling program [Program effective January 5, 2025]. https://www.mta.info/project/CBDTP

Needell, E., Amirgholy, V., & Asudegi, M. (2023). Estimating macroscopic fundamental diagrams for u.s. urban networks using probe-vehicle data and machine learning. *Transportation Research Part B: Methodological*, *158*, 99–115. https://doi.org/10.1016/j.trb.2023.08.003

Pigou, A. C. (1920). *The economics of welfare*. Macmillan; Co.

Rahman, M., Gurumurthy, K. M., & Kockelman, K. M. (2021). Impact of flextime on departure time choice for home-based commuting trips in austin, texas. *Transportation Research Record: Journal of the Transportation Research Board*, *2676*(1), 446–459. https://doi.org/10.1177/03611981211035757

Schrank, D., Albert, L., Jha, K., & Eisele, B. (2024, June). *2023 Urban Mobility Report* (tech. rep.). Texas A&M Transportation Institute, The Texas A&M University System. https://mobility.tamu.edu/umr/report/

Schrank, D., Eisele, B., & Lomax, T. (2019, August). *2019 urban mobility report* (tech. rep.) (Sponsored by the Texas Department of Transportation). Texas A&M Transportation Institute, Texas A&M University System. College Station, Texas. https://mobility.tamu.edu/umr/

Small, K. A. (1982). The Scheduling of Consumer Activities: Work Trips. *The American Economic Review*, *72*(3), 467–479.

Small, K. A. (2015). The bottleneck model: An assessment and interpretation. *Economics of Transportation*, *4*(1), 110–117. https://doi.org/10.1016/j.ecotra.2015.01.001

Small, K. A., & Verhoef, E. T. (2007). *The economics of urban transportation*. Routledge. https://doi.org/10.4324/9781315157375

Stockton, N. (2017). Why can't america have a single tolling system? because bureaucracy. *WIRED*. https://www.wired.com/story/national-tolling-system/

Takayama, Y. (2015). Bottleneck Congestion and Distribution of Work Start Times: The Economics of Staggered Work Hours Revisited. *Transportation Research Procedia*, *7*, 499–518. https://doi.org/10.1016/j.trpro.2015.06.026

Theseira, W. (2020). *Congestion control in singapore* (Discussion Paper No. 2020/10). International Transport Forum. Paris, OECD Publishing. https://doi.org/10.1787/7d266609-en

U.S. Department of Transportation. (2016, September). *Revised departmental guidance on valuation of travel time in economic analysis* (tech. rep.) (Issued September 27; Effective December 16, 2016). U.S. Department of Transportation, Office of the Secretary. Washington, DC. https://www.transportation.gov/sites/dot.gov/files/docs/2016%5C%20Revised%5C%20Value%5C%20of%5C%20Travel%5C%20Time%5C%20Guidance.pdf

Van Den Berg, V., & Verhoef, E. T. (2011). Winning or losing from dynamic bottleneck congestion pricing? *Journal of Public Economics*, *95*(7-8), 983–992. https://doi.org/10.1016/j.jpubeco.2010.12.003

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. https://doi.org/10.2307/1911814

Whitty, J. M., & Svadlenak, J. R. (2009). *Discerning the pathway to implementation of a national mileage-based charging system* (Special Report No. 299). Transportation Research Board of the National Academies. Washington, D.C. https://onlinepubs.trb.org/onlinepubs/sr/SR299Mileage.pdf

Wolff, H. (2014). Value of time: Speeding behavior and gasoline prices. *Journal of Environmental Economics and Management*, *67*(1), 71–88. https://doi.org/10.1016/j.jeem.2013.11.002

Yang, J., Purevjav, A.-O., & Li, S. (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. https://doi.org/10.1257/pol.20170195

Yu, X., Van Den Berg, V. A., & Verhoef, E. T. (2022). Autonomous cars and activity-based bottleneck model: How do in-vehicle activities determine aggregate travel patterns? *Transportation Research Part C: Emerging Technologies*, *139*, 103641. https://doi.org/10.1016/j.trc.2022.103641

Zhu, P. (2013). Telecommuting, household commute and location choice. *Urban Studies*, *50*(12), 2441–2459. https://doi.org/10.1177/0042098012474520