NBER WORKING PAPER SERIES

DOES HYPERCONGESTION EXIST? NEW EVIDENCE SUGGESTS NOT

Michael L. Anderson
Lucas W. Davis

Working Paper 24469
http://www.nber.org/papers/w24469

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2018

Neither of us have received any financial compensation for this project, nor do we have any financial relationships that relate to this research. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Hypercongestion Exist? New Evidence Suggests Not
Michael L. Anderson and Lucas W. Davis
NBER Working Paper No. 24469
March 2018
JEL No. C36,H23,R41,R42,R48

ABSTRACT

Transportation engineers are taught that as demand for travel goes up, this decreases not only speed but also the capacity of the road system, a phenomenon known as hypercongestion. We revisit this idea. There is no question that road systems experience periods in which capacity falls. However, we point out that capacity is determined by both demand and supply. Road construction, lane closures, stalled vehicles, weather, and other supply shocks provide an alternative explanation for the empirical evidence on hypercongestion. Using data from the Caldecott Tunnel in Oakland, California, we show that a naive regression recovers the standard hypercongestion result in the literature. However, once we use instrumental variables to isolate the effect of travel demand this effect disappears and across specifications we find no evidence that capacity decreases during periods of high demand. This lack of evidence of hypercongestion calls into question long-standing conventional wisdom held by transportation engineers and implies that efficient "Pigouvian" congestion taxes should be substantially lower than implied by hypercongestion models.

Michael L. Anderson
Department of Agricultural and Resource Economics
207 Giannini Hall, MC 3310
University of California, Berkeley
Berkeley, CA 94720
and NBER
mlanderson@berkeley.edu

Lucas W. Davis
Haas School of Business
University of California
Berkeley, CA 94720-1900
and NBER
ldavis@haas.berkeley.edu

1 Introduction

The relationship between the number of vehicles on the road and the speed at which they travel is fundamental to transportation and urban economics. To anyone who has driven in traffic, it is clear that traffic congestion decreases speed. But transportation engineers go further, arguing that as demand for travel goes up, this decreases not only speed but also the *capacity* of the road system, a phenomenon known as hypercongestion. Many studies claim to document what has been called "capacity drop", "flow breakdown", or the "two capacity phenomenon" (Banks, 1990, 1991; Hall and Agyemang-Duah, 1991; Bertini and Malik, 2004; Persaud et al., 1998; Cassidy and Bertini, 1999; Zhang and Levinson, 2004; Chung et al., 2007; Oh and Yeo, 2012).

Our paper revisits this idea. There is no question that road systems experience periods in which capacity falls. However, we point out that capacity is determined by both demand and supply. Road construction, accidents, lane closures, disabled vehicles, weather, and other supply shocks provide an alternative explanation for the empirical phenomenon of hypercongestion. We argue this is equivalent to the standard problem of estimating a supply curve using market data on prices and quantities. Economists have long used instrumental variables in these settings to econometrically separate demand and supply (see, e.g., Angrist and Krueger, 2001). We use the same approach here.

Our empirical analysis uses high-frequency traffic data from the Caldecott Tunnel in Oakland, California. This bottleneck is a good location for studying traffic congestion because traffic delays are common; indeed, transportation engineers have frequently studied this exact location (Chin and May, 1991; Chung and Cassidy, 2002; Chung et al., 2007). We observe traffic speed and vehicle counts at several locations before, during, and after the bottleneck. We first show that naive regressions can recover the standard hypercongestion result in the literature. However, once we use instruments to isolate the effect of travel demand, this relationship disappears, and across a variety of specifications we find no evidence that capacity decreases during periods

of high demand.

This absence of evidence of hypercongestion calls into question long-standing conventional wisdom in the transportation engineering literature. Hypercongestion has long been, for example, the primary rationale for freeway metering lights and other traffic interventions aimed at regulating demand (Diakaki et al., 2000; Smaragdis et al., 2004; Cassidy and Rudjanakanoknad, 2005), but our evidence raises questions about the applicability of these interventions. Our results also imply that the efficient "Pigouvian" congestion taxes should be substantially lower than implied by hypercongestion models. Using a stylized simulation model we show that without hypercongestion the marginal damages from driving are about half as large.

Previous studies by economists have tended to take hypercongestion as given. For example, Walters (1961) takes a supply curve with hypercongestion from the transportation engineering literature and uses the parameters to derive efficient congestion prices. Similarly, Keeler and Small (1977) uses a supply curve with hypercongestion to develop a long-run model of highway pricing and investment. A handbook chapter titled "Hypercongestion", Small and Chu (2003), explains that the standard "engineering relationship" has a backward-bending region known as hypercongestion, and then presents models with this feature. Most recently, Hall (forthcoming) uses a hypercongestion model to show how highway pricing can generate a Pareto improvement when agents are homogeneous, even before redistributing toll revenues.

Our paper is germane to a growing empirical literature on the economic costs of traffic congestion. Couture et al. (forthcoming) develops an econometric methodology for estimating city-level supply curves for trip travel, and constructs travel speed indices for large U.S. cities. Akbar and Duranton (2017) uses travel surveys and other data from Bogotá, Colombia to estimate the deadweight loss of traffic congestion. Adler et al. (2017) examines the effects of public transit strikes in Rome on arterial road congestion and concludes that hypercongestion, while rare, accounts for approximately 30 percent of congestion-related welfare losses.[1]

[1]Farther afield, there are also a number of studies by economists that examine the effect of building highways on traffic congestion, suburbanization, and other outcomes (see, e.g. Baum-

2 Conceptual Framework

Panel A of Figure 1 comes from the *Highway Capacity Manual*, the standard reference text in transportation engineering published by the Transportation Research Board, a division of the U.S. National Research Council (Transportation Research Board, 2016). The figure claims to show a supply curve for travel. With few vehicles on the road, travel is unconstrained, and average speeds are about 70 mph. Travel continues unconstrained at low demand levels, and then speeds slowly decrease until flow reaches a maximum capacity of about 2,000 vehicles per lane per hour. At these high demand levels, speeds are slower, and the "price" of travel (time spent) is higher.

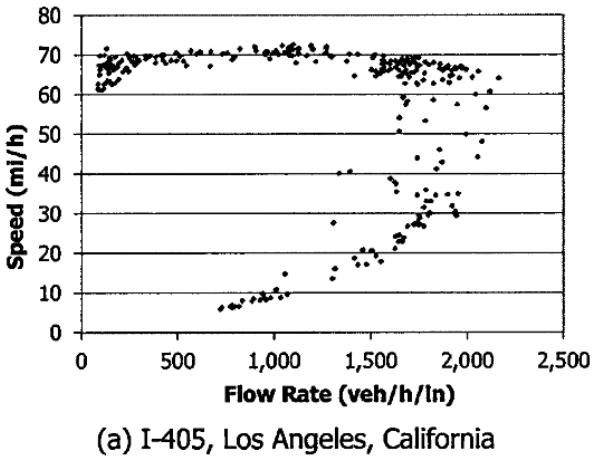
What is surprising, however, is that the curve then appears to bend backward. The manual explains that further increases in travel demand generate "flow breakdown", causing severe decreases in speed as well as decreases in freeway *capacity*, with flow rates falling well below the observed maximum. Transportation engineers are taught that this backward-bending supply curve is one of the "basic relationships" in traffic. The manual attributes these low-speed, low-flow observations to "oversatuated flow" arising from excess demand.

We have a simple alternative explanation. Without question there are periods in which the capacity of a road system decreases. But we point out that supply shocks provide an important alternative explanation for these low-speed, low-flow observations. Construction, lane closures, accidents, disabled vehicles, weather, and other supply shocks represent a "shift" in the supply curve, rather than a movement along the curve. Speed and flow are determined simultaneously through the interaction of demand and supply. Our argument is that because both demand and supply are shifting, data like the scatterplot above should be interpreted as a locus of possible equilibria, rather than as a supply curve.

Snow, 2007; Duranton and Turner, 2011). Another strand of the economics literature on traffic congestion considers "bottleneck" models in which drivers face a tradeoff between time delays and when to leave for a trip (Vickrey, 1969; Small, 1982; Arnott et al., 1990, 1994). Mun (1999), in particular, considers a bottleneck model with hypercongestion.

Figure 1: Travel Supply Curves?

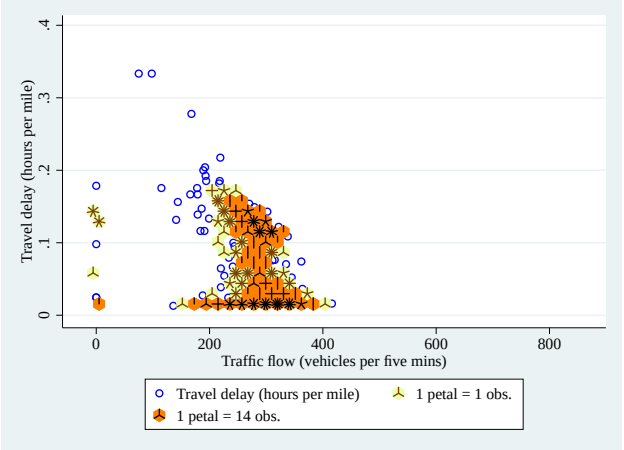
Panel A: Hypercongestion?



(a) I-405, Los Angeles, California

Note: Reprinted from *Highway Capacity Manual*, 2016, "Relationships Among Basic Parameters"

Panel B: Hypercongestion or Shifting Supply Curve?



Panel B of Figure 1 plots data on delays (the inverse of speed) and traffic flows using observations from our study location when the bottleneck is in effect. The basic pattern is very similar to Panel A; traffic delays tend to increase with traffic flow. However, there are also low-speed, low-flow observations which make the supply curve appear to bend backward. In the sections that follow, we show that these observations are the result of supply shocks. Using standard econometrics, and in particular, two stage least squares (2SLS) regressions, we empirically separate these "shifts" in the supply curve from movements along the curve. Our instrument is a demand shifter — a variable correlated with demand but not with supply — which allows us to hold supply factors fixed while tracing out the causal relationship between travel demand and freeway capacity.

Before proceeding, we note two important caveats. First, our study focuses on freeways, not arterial street networks. Freeways are a vital component of the road network, accounting for the majority of vehicle miles traveled (Lomax et al., 2018). Indeed, all of the transportation engineering papers that we cite above focus on freeways. Freeway geometry, however, differs fundamentally from arterial road geometry because freeways lack intersections with conflicting cross traffic. Our results do not speak to whether hypercongestion occurs on a dense street network with conflicting directions of traffic. Second, our setting is a standard bottleneck in which multiple lanes feed into a lesser number of lanes. In more esoteric settings, a queue from a bottleneck on one route may spill over onto a different route that does not traverse the bottleneck, creating a "triggerneck" (Vickrey, 1969). Our results do not apply to triggernecks.

3 Empirical Application

3.1 Caldecott Tunnel

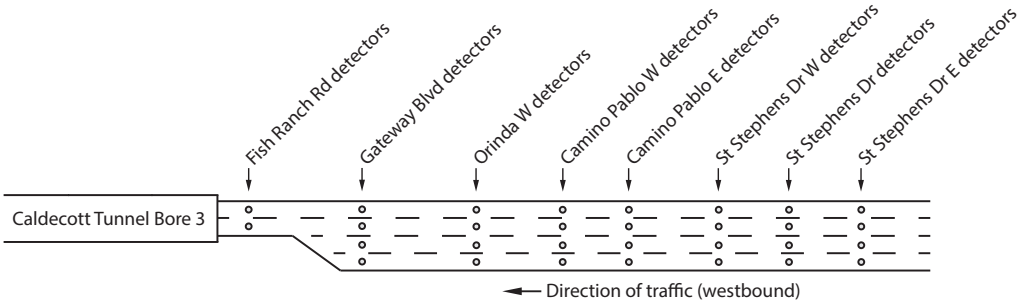
Our empirical application examines westbound traffic along California State Route 24 (SR-24) at the Caldecott Tunnel. SR-24 connects suburban Contra Costa County, to the east, with the urban cities of Oakland and San Francisco, to the west. This location is a classic bottleneck, with the number of lanes decreasing as traffic approaches the tunnel. During the study period the tunnel featured two reversible lanes that operated westbound in the morning and eastbound in the afternoon and evening. We focus on weekday afternoons and evenings from 2005 to 2010, a period and set of hours during which the Caldecott Tunnel was operated such that westbound vehicles had to merge from four lanes to two as they approached the tunnel.[2]

Figure 2 shows the study location. Approximately 3,000 feet before the tunnel the number of lanes merges from four down to two. This is the key feature of our study location and the place where the vehicle queue typically begins. The figure also shows using small circles the location of 30 "loop detectors", our primary source of data on both traffic flows and average vehicle speed.[3] We observe a set of two loop

[2]Rather than a single wide tunnel, the Caldecott consists of multiple "bores", each with two lanes carrying traffic in a single direction. Although the tunnel was expanded to four bores (eight total lanes) in 2013, we study the period from 2005 to 2010 when the tunnel still had only three bores and construction had not yet begun on the fourth bore. During the period we study, the middle bore was operated westward during morning hours as commuters drive toward Oakland and San Francisco, and eastward during afternoon and evening hours as commuters drive toward suburban Contra Costa County. We focus on weekday afternoons because this is when the middle bore was operated eastward. The middle bore also operated eastward during some weekend hours. However, we do not have reliable documentation of these hours, so our analysis focuses on weekdays. Afternoon westbound traffic is lighter than eastbound traffic, but with only a single bore open in that direction, the bottleneck was more than sufficient to generate significant traffic delays. Finally, we do not use the eastbound morning bottleneck in our analysis because it features traffic merging from multiple directions, making it impossible to measure queue length, which we use later as a proxy for demand.

[3]Loop detectors are small insulated electric circuits installed in the middle of traffic lane. Loop detectors measure the rate at which vehicles pass, e.g. vehicles per five-minute period. In addition,

Figure 2: Traffic Merge Heading into Caldecott Tunnel



detectors after the merge but before the tunnel (Fish Ranch), as well as a series of loop detectors upstream of the merge starting with Gateway, and followed by Orinda West, Camino Pablo West, Camino Pablo East, St. Stephens West, St. Stephens, and St. Stephens East.[4]

Before introducing our main empirical models we reproduce the hypercongestion and capacity drop results from the existing literature using several "naive" regressions. In our data there is a clear positive relationship between vehicle speed and traffic flows. For example, when we estimate a time series regression using observations at Gateway Blvd during weekday daytime hours, we find that a 50% reduction in speed correlates with a 4% drop in total traffic flow across all lanes ($t = 14.8$). This positive correlation between speed and traffic flow is typical of the type of evidence that has been used in the literature to demonstrate hypercongestion. In the context of our empirical application, however, this relationship is clearly an artifact of the

loop detectors measure average vehicle speed by sensing how long it takes each vehicle to pass over the detector. These loop detectors are maintained by the California Department of Transportation (Caltrans), and data are made publicly available through the *Performance Measurement System* (PeMS) at *http://pems.dot.ca.gov/*. Subsequent to our sample dates, the Gateway Blvd exit was renamed Wilder Rd.

[4]For visual clarity the figure does not include freeway exits and entrances. One of the significant advantages of the study location is that there are relatively few exits and entrances nearby. The last freeway entrance prior to the bottleneck is approximately 9,000 feet (1.7 miles) west of the tunnel, near the Orinda West detectors (there are also exits and entrances at Gateway Blvd, but they did not connect to any roads at the time). While it is possible for drivers to avoid the tunnel completely by taking the Fish Ranch Rd exit, the drive over the hill is slow and circuitous, and, regardless, the bottleneck has already been cleared by the time a driver reaches Fish Ranch Rd.

daily supply shock — reversible lane closures — that simultaneously lowers traffic flow and decreases speeds. This is a somewhat extreme case, but it illustrates how a shifting supply curve can easily generate a positive correlation between speed and traffic flows, causing it to appear as if the supply curve bends backwards.

Moreover, even restricting the sample to include only afternoon and evening hours when the bottleneck is always present, there is still a pronounced positive cross-sectional relationship between vehicle speeds and traffic flows. For this next "naive" regression, we take speed and traffic flow measurements from six lanes – two downstream (Fish Ranch) and four upstream (Gateway) of the bottleneck, generating six observations for each five-minute period. When we then regress traffic flows on speed while including fixed effects for every five-minute period in the sample (thus isolating cross-sectional variation), we find that a 50% reduction in speed correlates with a 21% decrease in traffic flows ($t = 48.3$). This relationship is nearly mechanical — each of the four lanes upstream of the bottleneck must process fewer vehicles per minute, at slower speeds, than each of the two lanes downstream of the bottleneck. These results are not causal relationships, but they highlight the ease with which an observer might conclude that hypercongestion and the capacity-drop phenomena are present in our data.

These "naive" regressions also highlight that traffic flows and capacity are not the same thing. Capacity is the maximum possible traffic flow, so it may differ significantly from actual traffic flows during off-peak periods. Accordingly, in the analyses that follow, we restrict the sample to include only observations from periods in which a queue has formed. By definition, if there is no queue, then traffic flows have not reached their maximum capacity, and measurements are limited by insufficient demand rather than maximum capacity. We refer to this condition as "demand starvation" — the bottleneck could process more vehicles were they available. Thus, in the regression analyses below we compare traffic flows between periods with different lengths of non-zero queues. We refer to traffic flows during these periods as "capacity", and ask whether, consistent with the capacity-drop hypothesis, increases in travel demand cause capacity to decrease.

3.2 Summary Statistics

The unit of observation in our data is a five-minute period. The sample includes all weekday observations during which a queue is present between 4 pm and 7 pm and June 2005 and June 2010 (Section 3.3 discusses the choice of 4 pm to 7 pm in more detail). We determine whether or not there is a queue and measure the length of the queue using the loop detectors upstream of the bottleneck. The first upstream loop detector, Gateway, is 1,690 feet before the bottleneck, and we observe approximately equidistant detectors all the way to St. Stephens East, which is 15,690 feet (about three miles) away from the bottleneck. For each loop detector we determine whether a queue is present by measuring whether traffic is moving at under 30 miles per hour (mph).[5] Thus, we define a queue as present if the average speed at Gateway is below 30 mph. The length of the queue is then measured using the number of consecutive upstream detectors for which we observe a delay. For example, if the measured speed is below 30 mph at the first detector (Gateway), but not at the second (Orinda West), then we conclude that the queue is 1,690 feet in length. "Broken" queues (e.g. a case in which traffic moves below 30 mph at Gateway, above 30 mph at Orinda West, and below 30 mph at Camino Pablo West) are rare in our estimation sample, and our results are robust to their inclusion or exclusion.

We measure speed and capacity downstream of the bottleneck using the loop detector at Fish Ranch Road. This loop detector is after traffic has merged from four to two lanes and cleared the bottleneck, but it is still east of the Caldecott Tunnel.[6] We measure speed upstream of the bottleneck using the loop detector at Gateway.

Appendix Table A1 reports summary statistics for our estimation sample.[7] It is

[5]This threshold is arbitrary, but the results are robust to alternative definitions. In Appendix Table A4 we use thresholds of 20 mph and 40 mph and find qualitatively similar results.

[6]In alternative specifications we use a location, Tunnel Road, which is west of the tunnel. However, the Tunnel Road data are not available prior to 2010, and in late 2010 construction began on the Caldecott fourth bore, causing unpredictable westbound delays near Tunnel Road over the next three years. The beginning of construction also coincided with the removal of the Fish Rand Road detector from service.

[7]Appendix Table A2 reports summary statistics for the full weekday sample, including periods with no queue.

worth making several observations about the summary statistics. First, as expected, average speed declines significantly at the bottleneck. Downstream of the bottleneck the average speed is over 40 mph, but upstream the average speed is only 11 mph. Speed downstream of bottleneck is still below the speed limit of 50 mph because the road is two lanes, so vehicles are still densely packed, and because the downstream loop detectors are close enough to the bottleneck that vehicles are still accelerating. Second, the average measured queue length is about 4,000 feet (0.8 miles), but the range is very wide, and at times the queue reaches as long as 15,700 feet. This maximum queue length is very long, about 3 miles, and speaks to the severity of the bottleneck. Third, there is also wide variation in capacity, ranging from zero to 408 vehicles per five minutes. The low-capacity observations are striking given that these are all periods with a non-zero queue, and, we argue, are due primarily to supply shocks (e.g. lane closures from accidents).[8]

3.3 Estimation Sample

Avoiding observations with demand starvation is a key statistical challenge — it is impossible to measure a bottleneck's capacity if it has not reached capacity. A simple strategy is to condition the estimation sample on the presence of a queue, regardless of time of day. This sample selection criterion, however, can introduce subtle but meaningful bias into an analysis that leverages exogenous variation in demand.

Consider periods of low demand. These periods typically have no queues, and if a queue is present it tends to be short. But if demand is low, why does a queue form at all? One possibility is a negative supply shock, such as an accident or poor weather. When conditioning on a queue existing, negative supply shocks will thus be more common during periods of low demand than during periods of high demand, as

[8]Observations with zero vehicles in a five-minute period account for less than 0.3% of observations in our main estimation sample. While there are valid reasons to observe zero vehicles, it is also possible that these values are measurement errors that the PeMS data-quality algorithms failed to catch. In Appendix Table A4 we verify that dropping these zero-vehicle observations has no impact on our estimates.

supply shocks are necessary for the formation of queues during periods of low demand. This happens even if supply shocks occur with equal frequency across all periods. The sample selection criterion itself can thus generate a positive relationship between capacity and queue length (our indicator of demand), even when both demand shocks and supply shocks are exogenous.

To mitigate this bias we focus on periods when demand should be sufficient to generate a queue even in the absence of a negative supply shock. Appendix Figure A1 plots the share of weekdays that a queue is present at each time of day. Visual inspection suggests that we focus on the period from 4 pm to 7 pm, which comprises our baseline estimation sample. To further narrow our sample, in some specifications we focus on times of day when there is at least a 75% chance of a queue (4:30 pm to 6:15 pm) and days of week with the highest frequency of early evening queues (Wednesday, Thursday, and Friday).[9]

In our most restrictive samples we further limit the sample to include only days predicted to be high demand using a machine-learning model. Specifically, we predict whether there is a continuous queue from 4:30 pm to 6:15 pm as a function of month-of-sample indicators, year-by-day-of-week indicators, calendar-month-by-day-of-week indicators, and third-order polynomials in hourly inflows (well upstream of the bottleneck) for each hour from 12 pm to 4 pm. In total we have 105 predictors.

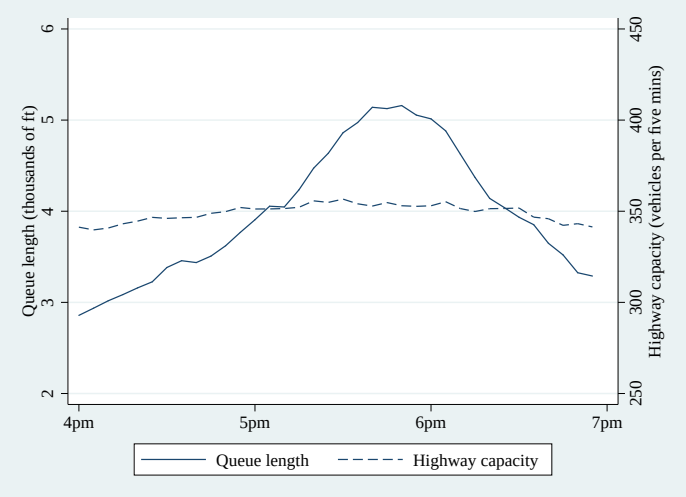
Given the large number of predictors, overfitting is a serious concern.[10] We thus estimate our regression using LASSO, which penalizes the model for each additional non-zero coefficient that it fits. The LASSO estimates non-zero coefficients for 17 predictors, the most important of which is, unsurprisingly, inflows at 4 pm.

Figure 3 presents a visual depiction of our empirical strategy using data from our

[9] At 5 pm, there is a 92% chance of observing a queue on a Wednesday, Thursday, or Friday. In comparison, there is only a 32% chance of a queue at 5 pm on Mondays, and a 76% chance of a queue at 5 pm on Tuesdays.

[10] With a sufficiently rich set of predictors, the regression could identify all days with continuous queues from 4:30 pm to 6:15 pm and drop all other days. This strategy risks including some days on which demand is relatively low but supply shocks occur.

Figure 3: Queue Length and Freeway Capacity



baseline estimation sample. The solid line plots average queue length (for non-zero queues) by time-of-day from 4 pm to 7 pm. The sharp increase in queue length in the early evening and the subsequent decrease later strongly suggest that demand varies with time-of-day. The solid line is thus visual evidence of a strong "first-stage" relationship between queue length and time-of-day. The dashed line plots average capacity (traffic flows) by time-of-day (again conditional on a non-zero queue). Its flat profile is evidence of no "reduced-form" relationship between capacity and time-of-day. Taken together the two lines suggest that increases in traffic demand are not accompanied by decreases capacity, calling into question the hypercongestion hypothesis.

3.4 Main Regression Estimates

Table 1 reports our main results. We show estimates and standard errors from four two stage least squares (2SLS) regressions of the following form,

$$capacity_t = \alpha + \beta \cdot queue\ length_t + \varepsilon_t. \tag{1}$$

The dependent variable in all regressions is capacity, measured downstream of the bottleneck in vehicles per five minutes. The independent variable of interest is the queue length, measured in thousands of feet. The coefficient of interest is β, which is the change in capacity associated with a one-unit increase in queue length. Motivated by Figure 3, we instrument for queue length using a third-order polynomial in time-of-day. Time-of-day is highly predictive of travel demand and the instrument F-statistics from the first-stage regression are large across the four columns — 61, 71, 87, and 23, respectively — indicating that we do not have a weak instruments problem.

We consider four different specifications with increasingly stringent criteria for selecting the sample. The main sample in Column (1) includes all five-minute periods on weekdays with a non-zero queue between 4 pm and 7 pm and June 2005 and June 2010. In Column (2) we restrict the sample to focus on Wednesdays through Fridays between 4:30 pm and 6:15 pm. We then further restrict the sample in Column (3) to include only days on which our LASSO model predicts a queue that persists from 4:30 pm to 6:15 pm with greater than 95% probability. Finally, in Column (4) we restrict the sample to summer months only, when traffic is least likely to be affected by adverse weather and lighting conditions.

The 2SLS procedure is equivalent to first regressing queue length on a third-order polynomial in time-of-day and then using the fitted values from that first-stage rather than queue length itself as the independent variable. Thus, the 2SLS regression is estimated using the predictable time-of-day variation in queue length rather than idiosyncratic day-to-day variation. This is desirable because the predictable time-of-day variation in queue length is mostly driven by differences in travel demand, while the day-to-day variation is also affected by construction delays, accidents, and other supply shocks.[11] The only supply shock evident to us that might correlate with

[11] An alternative to this instrumental variables strategy would be to exclude observations affected by supply shocks. This ends up being impractical, however, because not all supply shocks are observed. The California Department of Transportation's *Performance Measurement System (PeMS)* tracks incidents reported by the California Highway Patrol including accidents and vehicle breakdowns. These data provide an accurate record of major accidents, but many smaller incidents are

time-of-day is lighting conditions. To check the sensitivity of our results to lighting conditions or inclement we restrict the sample to summer months only in the last column. In these months sunset generally occurs after 8 pm, and total monthly precipitation never exceeds 0.1 inches.[12]

Across specifications, there is no evidence of hypercongestion. If hypercongestion were present, we would expect the estimates of β to be negative, with increased demand leading to a decrease in capacity. Instead, all four estimates of β are positive. In the specification in Column (1), each additional 1,000 feet of queue is associated with a 6.1 vehicle per five minute increase in capacity. This coefficient is small compared to mean capacity (350), but strongly statistically significant. Column (2) focuses on periods when demand is likely to be sufficient to generate a queue. The coefficient estimate drops to 2.7 and remains highly significant (but still modest in comparison to mean capacity). Further restricting the sample to only include periods when our LASSO model predicts a queue or during summer months — Columns (3) and (4) — reduces the coefficient estimates to 1.7 and 2.3 and eliminates statistical significance. These estimates are precise enough to rule out capacity drops of more than 0.3% from each additional 1,000 feet of queue.

3.5 Additional Specifications and Robustness Checks

In the online appendix we present results from several alternative specifications and robustness checks. First, when we estimate these regressions using least squares, rather than 2SLS, the estimates are much smaller, close to zero, and not statistically different from zero. Appendix Table A3 reports these results. This pattern is not

unreported. In addition, the PeMS incident data report the date of the incident, but not the exact time, making this less relevant for our high-frequency analysis. Moreover, there are lane closures, adverse weather conditions, and other types of supply shocks that are not recorded in PeMS.

[12]The median sunset from June to August at our location occurs at 8:26 pm, and the earliest sunset occurs around 7:38 pm. Furthermore, the hill through which the Caldecott Tunnel bores has a peak of 1,400 feet, which is at least 500 feet higher than the summit of SR-24. Thus drivers never experience sunlight shining directly in their eyes, as the sun gets blocked by the hill at least 90 minutes prior to sunset.

surprising given that queue length is driven by both demand and supply shocks.

Next we check whether large outliers, in either the positive or negative direction, can explain our results. We are specifically concerned about outliers in the negative direction, as these may represent accidents; if accidents occur disproportionately around 4 pm or 7 pm, when queues are shorter, they could bias our 2SLS estimates. To test whether outliers drive our results we estimate least absolute values (LAV) regressions instead of least squares regressions. LAV regressions estimate the conditional median of the dependent variable, whereas OLS estimates the conditional mean. Median regressions are, by definition, insensitive to outliers. Appendix Table A5 reports estimates from two conditional quantile instrumental variables regressions (analogous to 2SLS) in Columns (1) and (2) and two LAV regressions in Columns (3) and (4) (analogous to OLS). In all cases the coefficient estimates are close to zero and statistically insignificant, implying that additional demand has no negative effect on capacity.

In summary there is no evidence of hypercongestion in any specification. To the contrary, longer queues are associated with slightly higher capacity. This could be because the longer queue ensures that there is always another driver to fill small gaps during the merge, or because drivers exert more effort to merge quickly after waiting in a long queue, or perhaps even a composition effect in which the type of drivers who approach the tunnel at 6 pm are drivers who merge more quickly. Regardless of the mechanism, these regression estimates provide a striking lack of evidence of any hypercongestion.

4 Policy Implications

Thus, across specifications we find no evidence that capacity decreases during periods of high demand. This absence of evidence of hypercongestion calls into question long-standing conventional wisdom in the transportation engineering literature. In this section, we argue that this finding also has important policy implications. In

particular, we use a stylized simulation model to show how the presence (or lack) of hypercongestion affects the marginal damages from driving.

Traffic congestion is a negative externality (Pigou, 1920; Vickrey, 1963, 1969; Newbery, 1990). That is, when a motorist drives on a congested road, she decreases the average speed of all drivers on the road. Since drivers do not internalize the delays they impose on others, the efficient "Pigouvian" tax would tax drivers an amount equal to the value of the marginal delay they impose on other drivers.

We parameterize a stylized simulation model using data from the Caldecott Tunnel. Our objective is to describe traffic, queuing, and marginal damages of traffic congestion for a typical weekday afternoon and evening. As in the empirical analysis, we consider five-minute periods. We take travel demand as fixed and exogenous, and we use traffic inflows measured on a representative day well upstream of the bottleneck at the Orinda West loop detector.[13] The model determines queue lengths and delays for all drivers based on the maximum capacity of the bottleneck. For the simulation without hypercongestion, we set a maximum capacity of 352.2 vehicles per five minutes. This value calibrates the model so that the late afternoon queue starts around 3 pm, the maximum average queue length over a one-hour period reaches approximately 930 vehicles, and the queue dissipates around 7 pm. These features are consistent with the observed data for our representative day.

We then calculate marginal damages by introducing an additional 10 vehicles during a given hour of the day.[14] We calculate total delay (in minutes) both with and without this small perturbation, so the difference gives us marginal damages in minutes. For the simulation, we repeat this exercise for 4 pm, 5 pm, and 6 pm, and then calculate the average marginal damages in minutes. It is a large number. For example, each additional vehicle introduced at 5 pm increases delays for other vehicles by a total

[13] We choose the representative day to be as close as possible to the sample averages on three dimensions: fraction of time that a queue was present; average length of queue (when present); and average inflow of vehicles. This procedure selects February 7, 2006. However, our qualitative conclusions hold for any day on which there are nontrivial, but not interminable, queues.

[14] 10 vehicles corresponds to less than 10 seconds of traffic flow during normal conditions. We get similar results (per vehicle) if we introduce 100 vehicles, or just one vehicle.

of 149 minutes.

To simulate hypercongestion we rely on the literature that has tended to find a "capacity drop" of approximately 10% during periods of high demand (see the list of studies in the introduction as well as those summarized by Hall (forthcoming)). In particular, we assume that capacity during each hour is approximately 366.3 multiplied by $1 - (queue/9,000)$, where *queue* is measured in number of vehicles. This formulation implies that capacity drops linearly in the size of the queue. The maximum queue size that the model, so parameterized, generates over a one-hour period is about 1,000 vehicles, which reduces the capacity to 366.3 * (0.89) = 326 vehicles per five minutes. The "intercept" of approximately 366.3 generates similar average queue length as a model with a capacity of 352.2 and no capacity drop (our baseline model).[15] The comparison of the two models thus addresses the question of how introducing hypercongestion affects marginal damages while holding other features of the model approximately fixed.

Table 2 summarizes the results. In the scenario described above, in which the maximum capacity drop reaches just above 10% with hypercongestion, marginal damages without hypercongestion are 53% smaller than marginal damages with hypercongestion.[16] In a scenario in which the capacity drop reaches 15% with hypercongestion — a value that several transportation engineering studies have found — marginal damages without hypercongestion are 67% smaller than marginal damages with hypercongestion. That is, marginal damages are only about one-third as large. These stylized simulations could be refined further. For example, one could con-

[15]To equate average queue length over the day, the intercept must be 366.3. If we instead wish to equate maximum queue length over a one-hour period, the intercept must be 367.6, and if we wish to equate queue length at 5 pm, the intercept must be 359.9. All of these parameterizations generate qualitatively similar conclusions, however; in all three cases the marginal damages without hypercongestion are half as large or less as marginal damages calculated under hypercongestion.

[16]Inserting additional vehicles at 4 pm instead of 5 pm increases the share of marginal damages generated by hypercongestion from 53% to 64%. Inserting additional vehicles at 6 pm instead of 5 pm reduces the share of marginal damages generated by hypercongestion from 53% to 42%. Inserting additional vehicles earlier in the queue increases the total number of vehicles affected (i.e. the total number of vehicles that follow the marginal vehicle), and inserting them later in the queue decreases the total number of vehicles affected.

sider whether drivers at the tail end of the queue might in the long run change their departure times, which could modestly reduce damages.[17] The simulation nevertheless illustrates that welfare impacts depend strongly on wehther hypercongestion exists.

5 Conclusion

This paper revisits the idea of hypercongestion. When we naively examine average speeds and traffic flows we find evidence consistent with a large literature on hypercongestion. But once we use standard econometric techniques to disentangle demand and supply, this backward-bending supply curve disappears. Across a variety of specifications we find no evidence of hypercongestion.

How can this be? To anyone who has been stuck in heavy traffic, it certainly feels as if the capacity of the roadway is being restricted in these moments. We suspect, however, that this feeling is largely about speed rather than capacity. There is no question that as more vehicles crowd onto the road, speed decreases. But speed and capacity are not identical. Speed is readily apparent to drivers, but capacity requires careful measurement.

In addition, on freeways the feeling of being trapped in heavy traffic usually occurs in a queue, waiting to pass a bottleneck. By definition the capacity *per lane* when approaching a bottleneck must drop, as the number of lanes drops at the bottleneck. Nevertheless, we find that the capacity of the bottleneck itself — the rate at which vehicles pass through the bottleneck — does not drop with the length of the queue. In short, no matter how much travel demand increases, the number of vehicles passing through the bottleneck does not decrease.

[17] With optimal tolling the difference in marginal damages would presumably narrow since, by definition, there would rarely be any congestion and thus no capacity drop. The optimal tolling scenario is of limited policy relevance, however, since optimal tolling — or even suboptimal tolling — is rare in the real world.

Our findings imply that congestion taxes should be much lower than implied by hypercongestion models and that tolling or other demand management strategies, such as ramp meters, are unlikely to increase the capacity of the freeway system. Nevertheless, congestion taxes should not be zero. To the contrary, our simulation estimates show that, even without hypercongestion, the marginal damages from traffic congestion can be very large. Starting from zero — the level at which most roadways are currently taxed — leaves considerable headroom for increases.

References

Adler, Martin W., Federica Liberini, Antonio Russo, and Jos N. van Ommeren, "Road Congestion and Public Transit," *ITEA Conference Working Paper*, 2017.

Akbar, Prottoy and Gilles Duranton, "Measuring the Cost of Congestion in Highly Congested City: Bogotá," *University of Pennsylvania Working Paper*, 2017.

Angrist, Joshua D and Alan B Krueger, "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 2001, *15* (4), 69–85.

Arnott, Richard, Andre De Palma, and Robin Lindsey, "Economics of a Bottleneck," *Journal of Urban Economics*, 1990, *27* (1), 111–130.

_ , André De Palma, and Robin Lindsey, "The Welfare Effects of Congestion Tolls with Heterogeneous Commuters," *Journal of Transport Economics and Policy*, 1994, pp. 139–161.

Banks, James H, "Flow Processes at a Freeway Bottlenecks," *Transportation Research Record*, 1990, (1287).

_ , "Two-Capacity Phenomenon at Freeway Bottlenecks: A Basis for Ramp Metering?," *Transportation Research Record*, 1991, (1320).

Baum-Snow, Nathaniel, "Did Highways Cause Suburbanization?," *Quarterly Journal of Economics*, 2007, *122* (2), 775–805.

Bertini, Robert and Shazia Malik, "Observed Dynamic Traffic Features on Freeway Section with Merges and Diverges," *Transportation Research Record*, 2004, (1867), 25–35.

Cassidy, Michael J and Jittichai Rudjanakanoknad, "Increasing the Capacity of an Isolated Merge by Metering its On-Ramp," *Transportation Research Part B*, 2005, *39* (10), 896–913.

_ and Robert L Bertini, "Some Traffic Features at Freeway Bottlenecks," *Transportation Research Part B*, 1999, *33* (1), 25–42.

Chin, Hong C and Adolf D May, "Examination of the Speed-Flow Relationship at the Caldecott Tunnel," *Transportation Research Record*, 1991, *1320*, 75–82.

Chung, Koohong and Michael Cassidy, "Testing Daganzo's Behavioral Theory for Multi-lane Freeway Traffic," *California Partners for Advanced Transit and Highways (PATH)*, 2002.

_ , Jittichai Rudjanakanoknad, and Michael J Cassidy, "Relation Between Traffic Density and Capacity Drop at Three Freeway Bottlenecks," *Transportation Research Part B*, 2007, *41* (1), 82–95.

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

Diakaki, Christina, Markos Papageorgiou, and Tom McLean, "Integrated Traffic-Responsive Urban Corridor Control Strategy in Glasgow, Scotland," *Transportation Research Record*, 2000, (1727), 101–111.

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

Hall, Fred L and Kwaku Agyemang-Duah, "Freeway Capacity Drop and the Definition of Capacity," *Transportation Research Record*, 1991, (1320).

Hall, Jonathan D., "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways," *Journal of Public Economics*, forthcoming.

il Mun, Se, "Peak-Load Pricing of a Bottleneck with Traffic Jam," *Journal of Urban Economics*, 1999, *46* (3), 323–349.

Keeler, Theodore E and Kenneth A Small, "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways," *Journal of Political Economy*, 1977, *85* (1), 1–25.

Lomax, Tim, David Schrank, and Bill Eisele, "Congestion Data for Your City — Urban Mobility Information," 2018.

Newbery, David M, "Pricing and Congestion: Economic Principles Relevant to Pricing Roads," *Oxford Review of Economic Policy*, 1990, *6* (2), 22–38.

Oh, Simon and Hwasoo Yeo, "Estimation of Capacity Drop in Highway Merging Sections," *Transportation Research Record*, 2012, (2286), 111–121.

Persaud, Bhagwant, Sam Yagar, and Russel Brownlee, "Exploration of the Breakdown Phenomenon in Freeway Traffic," *Transportation Research Record*, 1998, (1634), 64–69.

Pigou, Arthur Cecil, *The Economics of Welfare*, London: Macmillan, 1920.

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

_ **and Xuehao Chu**, "Hypercongestion," *Journal of Transport Economics and Policy*, 2003, *37* (3), 319–352.

Smaragdis, Emmanouil, Markos Papageorgiou, and Elias Kosmatopoulos, "A Flow-Maximizing Adaptive Local Ramp Metering Strategy," *Transportation Research Part B: Methodological*, 2004, *38* (3), 251–270.

Transportation Research Board, "Highway Capacity Manual 6th Edition: A Guide for Multimodal Mobility Analysis," 2016.

Vickrey, William S, "Pricing in Urban and Suburban Transport," *American Economic Review*, 1963, *53* (2), 452–465.

_ , "Congestion Theory and Transport Investment," *American Economic Review*, 1969, *59* (2), 251–260.

Walters, Alan A, "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, 1961, pp. 676–699.

Zhang, Lei and David Levinson, "Some Properties of Flows at Freeway Bottlenecks," *Transportation Research Record*, 2004, (1883), 122–131.

Table 1: The Effect of Travel Demand on Freeway Capacity, 2SLS

VARIABLES	(1) Capacity	(2) Capacity	(3) Capacity	(4) Capacity
Queue length (1000s of feet)	6.051***	2.665***	1.212	2.253
	(0.712)	(0.714)	(0.675)	(1.201)
Observations	11,594	5,213	2,936	732
Dependent variable mean	349.7	353.5	357.1	359.4
Days	411	251	137	34
Days of week	Weekdays	Wed - Fri	Wed - Fri	Wed - Fri
Time of day	4 - 7pm	4:30 - 6:15pm	4:30 - 6:15pm	4:30 - 6:15pm
Require that queue predicted	No	No	Yes	Yes
Months	All	All	All	Jun - Aug

Notes: This table reports estimates and standard errors from four separate regressions, all estimated using two stage least squares (2SLS) with a third-order polynomial in time-of-day as the instrumental variables. The dependent variable in all regressions is freeway capacity, measured in vehicles per five minutes downstream of the bottleneck at Fish Ranch Road. The sample includes five-minute periods between June 2005 and June 2010 during which there is a queue. "Queue predicted" implies that the LASSO model predicted greater than 95% probability of a continuous queue from 4:30 to 6:15 pm. Queue length is measured in thousands of feet. Standard errors are clustered by day.

Table 2: Simulation Model Results

	Marginal damages w/ hypercongestion	Marginal damages w/o hypercongestion	Percent reduction
Baseline model w/10% capacity drop	315 mins	149 mins	53%
Alternative model w/15% capacity drop	449 mins	149 mins	67%

Notes: This table reports the results from our stylized simulation model of the Caldecott Tunnel. The first and second columns report the marginal damages in minutes imposed by an additional vehicle during weekday peak hours. In the first column the capacity of the bottleneck is assumed to be fixed and constant, whereas in the second column the capacity is assumed to drop by up to 10% or 15% due to hypercongestion. Finally, the last column reports the percent reduction in damages without hypercongestion. See the text for details.

Figure A1: Queue Presence by Time of Day

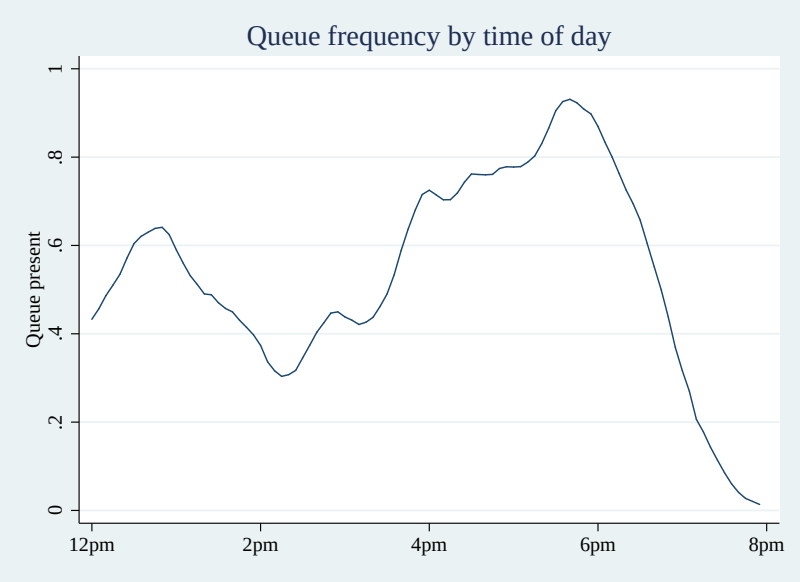


Table A1: Summary Statistics

Variable	N	Mean	Std Dev	Min	Max
Year	11,594	2008	1.800	2005	2010
Queue length (1000s of feet)	11,594	4.064	2.991	1.690	15.68
Capacity downstream of bottleneck (vehicles per 5 mins)	11,594	349.7	26.17	0	408
Speed upstream of bottleneck (mph)	11,594	11.00	6.643	3	45.90
Speed downstream of bottleneck (mph)	11,594	41.16	4.693	5.850	54

Notes: The sample includes all five-minute weekday periods between 4 pm and 7 pm and June 2005 and June 2010 during which a queue has formed. We define a queue as present at a detector during a five-minute period if the average speed over that detector falls below 30 mph. Speed upstream of the bottleneck is measured at Gateway Boulevard, averaged across four lanes. Speed downstream of the bottleneck is measured at Fish Ranch Road, averaged across two lanes.

Table A2: Summary Statistics

Variable	N	Mean	Std Dev	Min	Max
Year	272,571	2008	1.509	2005	2010
Queue present	272,571	0.088	0.284	0	1
Queue length (1000s of feet)	272,571	0.294	1.251	0	15.68
Capacity downstream of bottleneck (vehicles per 5 mins)	272,571	195.2	114.7	0	408
Speed upstream of bottleneck (mph)	267,205	59.34	16.22	3	76.28
Speed downstream of bottleneck (mph)	272,571	57.13	8.661	4.200	75.10

Notes: The sample includes all five-minute weekday periods with non-missing data between June 2005 and June 2010. We define a queue as present at a detector during a five-minute period if the average speed over that detector falls below 30 mph. Speed upstream of the bottleneck is measured at Gateway Boulevard, averaged across four lanes. Speed downstream of the bottleneck is measured at Fish Ranch Road, averaged across two lanes.

Table A3: The Effect of Travel Demand on Capacity, Least Squares

VARIABLES	(1) Capacity	(2) Capacity	(3) Capacity	(4) Capacity
Queue length (1000s of feet)	-0.435	-0.696	-0.203	0.351
	(0.411)	(0.374)	(0.282)	(0.432)
Observations	11,594	5,213	2,936	732
Dependent variable mean	349.7	353.5	357.1	359.4
Days	411	251	137	34
Days of week	Weekdays	Wed - Fri	Wed - Fri	Wed - Fri
Time of day	4 - 7pm	4:30 - 6:15pm	4:30 - 6:15pm	4:30 - 6:15pm
Require that queue predicted	No	No	Yes	Yes
Months	All	All	All	Jun - Aug

Notes: This table reports estimates and standard errors from four separate least squares regressions. The dependent variable in all regressions is freeway capacity, measured in vehicles per five minutes downstream of the bottleneck at Fish Ranch Road. The sample includes five-minute periods between June 2005 and June 2010 during which there is a queue. "Queue predicted" implies that the LASSO model predicted greater than 95% probability of a continuous queue from 4:30 to 6:15 pm. Queue length is measured in thousands of feet. Standard errors are clustered by day.

Table A4: Robustness Checks

VARIABLES	(1) Capacity	(2) Capacity	(3) Capacity	(4) Capacity	(5) Capacity	(6) Capacity
Queue length (1000s of feet)	2.744***	-0.763*	2.594***	-0.625	2.580***	-0.720
	(0.736)	(0.387)	(0.698)	(0.366)	(0.710)	(0.373)
Observations	5,164	5,164	5,261	5,261	5,200	5,200
Dependent variable mean	353.6	353.6	353.4	353.4	353.6	353.6
Days	249	249	252	252	250	250
Model	2SLS	OLS	2SLS	OLS	2SLS	OLS
Drop zero-flow observations	No	No	No	No	Yes	Yes
Queue speed threshold	20 mph	20 mph	40 mph	40 mph	30 mph	30 mph

Notes: This table reports estimates and standard errors from six separate regressions, estimated using either least squares or two stage least squares (2SLS) with a third-order polynomial in time-of-day as the instrumental variables. The dependent variable in all regressions is freeway capacity, measured in vehicles per five minutes downstream of the bottleneck at Fish Ranch Road. The sample includes five-minute periods between 4:30 pm and 6:15 pm, Wednesday and Friday, and June 2005 and June 2010 during which there is a queue (identical to Columns (2) in Tables 1 and A3). Queue length is measured in thousands of feet. Standard errors are clustered by day.

Table A5: The Effect of Travel Demand on Capacity, Median Regression

VARIABLES	(1) Capacity	(2) Capacity	(3) Capacity	(4) Capacity
Queue length (1000s of feet)	0.691	0.952	-0.162	-0.324
	(0.736)	(0.681)	(0.267)	(0.249)
Observations	5,213	2,936	5,213	2,936
Dependent variable mean	353.5	357.1	353.5	357.1
Days	251	137	251	137
Model	Median IV	Median IV	Median Reg	Median Reg
Require that queue predicted	No	Yes	No	Yes

Notes: This table reports estimates and standard errors from four separate median regressions, estimated using least absolute values (Median Reg) or conditional quantile instrumental variables (Median IV) with a third-order polynomial in time-of-day as the instrumental variables. The dependent variable in all regressions is freeway capacity, measured in vehicles per five minutes downstream of the bottleneck at Fish Ranch Road. The sample includes five-minute periods between 4:30 pm and 6:15 pm, Wednesday and Friday, and June 2005 and June 2010 during which there is a queue (identical to Columns (2) in Tables 1 and A3), with additional restrictions in Columns (2) and (4) (which are analogous to Columns (3) in Tables 1 and A3)). Queue length is measured in thousands of feet. Standard errors are clustered by day.