NBER WORKING PAPER SERIES

TWO EMPIRICAL TESTS OF HYPERCONGESTION

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, Revised May 2019

Previously circulated as "Does Hypercongestion Exist? New Evidence Suggests Not." Neither of us have received any financial compensation for this project, nor do we have any financial relationships that relate to this research. We are grateful to Gilles Duranton, Jonathan Hughes, Mark Jacobsen, Ian Parry, and Kenneth Small, as well as to the editor (Hunt Allcott) and three anonymous reviewers for helpful comments. 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.

Two Empirical Tests of Hypercongestion
Michael L. Anderson and Lucas W. Davis
NBER Working Paper No. 24469
March 2018, Revised May 2019
JEL No. C36,H23,R41,R42,R48

ABSTRACT

There is a widely-held view 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. We propose two empirical tests motivated by previous analytical models of hypercongestion. Our first test uses instrumental variables to empirically isolate the effect of travel demand on highway capacity. Our second test uses an event study analysis to measure changes in highway capacity at the onset of queue formation. We apply these tests to three highway bottlenecks in California for which detailed data on traffic flows and vehicles speeds are available. Neither test shows evidence of a reduction in highway capacity at any site during periods of high demand. Across sites and specifications we have sufficient statistical power to rule out small reductions in highway capacity. This lack of evidence of hypercongestion has important implications for travel supply and demand models and raises questions about highway metering lights and other traffic interventions aimed at regulating demand.

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 there is also a view 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. Standard reference texts show a backward-bending speed-flow curve (May, 1990; Lindsey and Verhoef, 2008; Transportation Review Board, 2016), and a number of economic analyses have taken hypercongestion as given (see, e.g. Walters, 1961; Johnson, 1964; Newbery, 1989; Mun, 1999; Fosgerau and Small, 2013; Hall, 2018a,b).

Our paper revisits this idea. We propose two empirical tests of hypercongestion. Both tests are novel in the literature but correspond closely with how previous studies have modeled hypercongestion. We apply our tests to three highway bottlenecks in California. We observe traffic flows and vehicle speeds at several locations before and after each bottleneck. Although the three sites differ, all have bottlenecks that generate slowed traffic and queues during afternoon hours.

Our first test uses instrumental variables to measure the effect of queue length on highway capacity. Though we are unaware of any previous attempt to estimate a regression of this form, our estimating equation corresponds closely with how previous studies have modeled hypercongestion (Yang and Huang, 1997; Fosgerau and Small, 2013; Small, 2015). We instrument using time-of-day to empirically isolate the effect of travel demand and find no evidence that capacity decreases during periods of high

demand. In short, no matter how long the queue gets, the number of vehicles passing through the bottleneck per minute remains stable.

Our second test uses an event study analysis to measure changes in traffic flows at the moment each afternoon when the queue initially forms. Several studies find what the literature calls "capacity drop", "flow breakdown", or the "two capacity phenomenon" at bottlenecks (Banks, 1990, 1991; Hall and Agyemang-Duah, 1991; Persaud et al., 1998; Cassidy and Bertini, 1999; Bertini and Malik, 2004; Zhang and Levinson, 2004; Chung et al., 2007; Oh and Yeo, 2012), referring to a drop in roadway capacity upon queue formation. In contrast to the vast majority of these studies, we find no evidence of a decrease in traffic flows at the moment the queue forms. Across sites, flows are essentially constant throughout the period of queue formation, with no discernible capacity drop whatsoever.

Thus we find no evidence of hypercongestion with either test. Though the findings are similar, the two tests are quite different, exploiting independent sources of variation and relying on distinct identifying assumptions. The results of the two tests are complementary, with neither test revealing a reduction in highway capacity as demand increases. This lack of evidence of hypercongestion is not due to a lack of statistical precision. At each site our data include tens of thousands of five-minute periods across hundreds of days. Given the modest fluctuations in observed flows during peak periods, this size of data set yields sufficient statistical power to rule out even small reductions in highway capacity. Throughout the analysis we report standard errors and 95% confidence intervals and show that we can reject economically significant capacity reductions, including those of the magnitudes suggested in the

existing literature.

These findings have significant policy implications. Hypercongestion has long been a rationale for highway ramp 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 marginal damages from driving are substantially lower than those implied by supply curves exhibiting hypercongestion. Nevertheless, as we explain in the paper, the implications for efficient "Pigouvian" congestion pricing are less clear.

Our paper is germane to a growing empirical literature on the formation of traffic congestion. Couture et al. (2018) develops an econometric methodology for estimating city-level supply curves for trip travel, and constructs travel speed indices for large U.S. cities. Yang et al. (2018) measures the effect of traffic density on speed in Beijing, using variation in traffic from driving restrictions. 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-Snow, 2007; Duranton and Turner, 2011). In other related work, Hanna et al. (2017) shows that elimination of high-occupancy vehicle lanes in Jakarta worsened traffic and Kreindler (2018) uses data from a smartphone app to study traffic congestion in Bangalore, India, finding at the city-level an approximately linear relationship between traffic volume and travel time.

Before proceeding, we note two important caveats. First, our study focuses on highways, not arterial street networks. Highways 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 highways. Highway geometry, however, differs fundamentally from arterial road geometry because highways lack conflicting cross traffic. Our results do not speak to whether hypercongestion occurs on a dense street network with conflicting directions of traffic. Second, our study focuses on standard bottlenecks in which the queue does not obstruct other upstream routes. Particularly in dense urban networks, a queue from a bottleneck on one route may sometimes spill over onto a different route that does not traverse the bottleneck, blocking that route and creating a "triggerneck" (Vickrey, 1969). Our results do not apply to triggernecks.

2 Conceptual Framework

2.1 Interpretations of the Speed-Flow Curve

Figure 1 shows eight examples of speed-flow curves. The first panel comes from the *Highway Capacity Manual*, the standard reference text in transportation engineering (Transportation Research Board, 2016), and the other panels come from the transportation literature (Drake and Schofer, 1966; Allen et al., 1985; May, 1990; Ni, 2015) and the economics literature (Keeler and Small, 1977; Newbery, 1989; Mun, 1999). In all cases, the horizontal axis measures traffic flow and the vertical axis measures

speed.

The upper part of the speed-flow curve exhibits a negative correlation between speed and flow. Speeds are high at low flow levels. For example, in the first panel, average speeds are near 70 miles-per-hour for flow levels below 1,500 vehicles/hour/lane. In all eight panels speeds are then lower at higher flow levels. In the first panel, for example, speeds tend to be between 60 and 70 miles-per-hour for flows above 1,500.

The lower part of the curve is more surprising, however. In all eight panels, there is a lower part of the curve that exhibits a *positive* correlation between speed and flow. Particularly striking are the observations with both very low speeds and very low flows. For example, the first panel from the *Highway Capacity Manual* includes observations with speeds below 10 miles-per-hour and flow rates below 1,000 vehicles/hour/lane.

There is a lack of consensus over how to interpret this lower region of the speed-flow curve. The *Highway Capacity Manual* explains that this region of the speed-flow curve exhibits "flow breakdown" and "oversaturated flow", with severe decreases in speed as well as decreases in capacity, and flow rates falling well below the observed maximum. This backward-bending curve is described as one of the "basic relationships" in traffic, with the lower part of the curve often drawn all the way to the origin, at which point both speed and flow are equal to zero.

Early economic analyses interpreted this speed-flow curve as a causal relationship. Walters (1961) and Johnson (1964), for example, interpreted the relationship as a

supply curve for travel, and used parametrized versions to derive efficient congestion prices. Later economic analyses similarly adopted a causal interpretation. For example, Newbery (1989) writes, "there is a sharp discontinuity in the mode of traffic flow as traffic reaches capacity. A small increase in the traffic flow, typically resulting from an extra inflow of traffic from a junction, causes a transition from relatively free flow to congested flow; speed drops sharply, and so does total flow." (p. 193).

Some later studies took a different, more descriptive interpretation. It is an empirical fact that low-speed, low-flow observations exist, but more recent economic analyses have argued that this relationship cannot be interpreted as a supply curve. For example, Small and Chu (2003) argues that "hypercongestion occurs as a result of transient demand surges and can be fully analyzed only within a dynamic model."[2] Similarly, Lindsey and Verhoef (2008) explains that these low-speed, low-flow observations occur "in queues upstream of a bottleneck".[3,4]

[2]This article, titled "Hypercongestion", notes that the standard "engineering relationship" has a backward-bending region known as hypercongestion. It then presents a series of dynamic models for straight uniform highways and dense street networks in which transient demand surges cause long vehicle queues, resulting in large travel time increases. It stresses the importance of studying hypercongestion using dynamic models, "Hypercongestion is a real phenomenon, potentially creating inefficiencies and imposing considerable costs. However, it cannot be understood within a steady-state analysis because it does not in practice persist as a steady state." (p. 342).

[3]While the term "hypercongestion" appears frequently in the literature, there is confusion regarding its exact definition. Mun (1999), for example, notes, "There is no commonly used terminology to represent [traffic jams] among disciplines. Although traffic engineers use the term 'congested flow,' this is not appropriate here because 'uncongested flow' in engineering is still regarded as congested flow in an economic sense. On the other hand, some economists use the term 'hypercongestion'..." (p. 323). Small and Chu (2003) defines hypercongestion as the region in the speed-flow diagram "in which speed increases with flow" and notes that it is "unsuitable as a supply curve for equilibrium analysis", (p. 319).

[4]Low-speed, low-flow observations are common upstream of bottlenecks. For example, Hall and Agyemang-Duah (1991) documents low-flow observations immediately upstream of a major highway entrance ramp. Relatedly, Sugiyama et al. (2008) and Tadaki et al. (2013) performed a pair of remarkable field experiments in which college students drove vehicles around a circle in

With or without explicit reference to the speed-flow curve, other later studies have continued to assume that there is a backward-bending part of the supply curve, i.e. that road capacity decreases at high levels of demand. Mun (1999), for example, constructs a bottleneck model with hypercongestion to calculate the gains from efficient congestion pricing.[5] Fosgerau and Small (2013) constructs a model in which road capacity declines with the length of the queue, and then uses this framework to calculate marginal damages and analyze alternative congestion pricing regimes. Hall (2018a) uses a hypercongestion model to show how highway pricing can generate a Pareto improvement when agents are heterogeneous, even before redistributing toll revenues.

2.2 Two Empirical Tests of Hypercongestion

In this paper we design and implement two empirical tests of whether highway capacity decreases during periods of high demand. To our knowledge, both tests are novel to this context, but both tests correspond closely with hypercongestion models proposed in the existing literature. Our first test uses instrumental variables to

an outdoor area and indoor baseball field, respectively. Varying the number of vehicles driving in the loop, the researchers demonstrate a pronounced decrease in vehicle flows as vehicle density increases. While they interpret this as evidence of low-speed, low-flow observations even without a bottleneck, an alternative interpretation would be that the loop effectively simulates the experience of being permanently in a queue, as the loop never empties into an uncongested "drain".

[5]In bottleneck models drivers face a tradeoff between time delays and trip departure times (Vickrey, 1969; Small, 1982; Arnott et al., 1990, 1993, 1994). Arnott (2013) argues that a key weakness of the standard bottleneck model is that it fails to incorporate hypercongestion, "In assuming that the discharge rate of the bottleneck is independent of the length of the queue behind it, the model assumes away hypercongestion, which most urban transport economists believe to be quantitatively important." (p. 119). Arnott (2013) proposes a "bathtub" model of hypercongestion for downtown areas in which capacity decreases at high levels of traffic density.

measure the effect of queue length on highway capacity. Our second test uses an event study analysis, measuring the change in traffic flows when a queue forms. The two tests use different sources of identifying variation and have different identifying assumptions.

Both tests of hypercongestion are designed to be applied in highway settings with a single bottleneck — locations where some physical feature of the highway serves to reduce traffic flow during periods of high demand. The most lucid example, and one that directly evokes the idea of the "neck" of a bottle, is a setting in which there is a sharp decrease in the number of lanes available for travel. We do not envision applying these tests to roadways with no spatial variation in capacity, which tend to have far fewer delays, or to dense urban road networks, which tend to have multiple sequential bottlenecks and alternative routing opportunities. As we noted earlier, highways account for a majority of vehicle miles traveled, and there is a large existing literature on highway bottlenecks, so our setting is an important one for understanding hypercongestion.

2.2.1 Instrumental Variables

Our first test takes the form of two stage least squares (2SLS) regressions of the following form,

$$\text{capacity}_t = \alpha_0 + \alpha_1 \cdot \text{queue length}_t + \varepsilon_t. \tag{1}$$

The dependent variable in these regressions is highway capacity in 5-minute period t, measured downstream of the bottleneck. The independent variable of interest

is queue length. The coefficient of interest is α_1, which is the change in capacity associated with a one-unit increase in queue length. We instrument for queue length using time-of-day. 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 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 weather, road construction, roadway incidents, and other supply shocks.[6]

Though we are not aware of any previous attempt to estimate a regression of this form, its basic structure coincides with several previous efforts to model hypercongestion. Small (2015) explains that hypercongestion can be modeled by "postulating that bottleneck capacity varies inversely with the length of the queue" (p. 115), citing several previous papers that have followed this approach. For example, Yang and Huang (1997) numerically solves a model in which capacity decreases exponentially with the length of the queue. Another example is Fosgerau and Small (2013). In their model, a bottleneck operates at full capacity for short queues, then drops discontinuously to a medium capacity when queue length exceeds a particular threshold, and then finally drops to zero capacity for very long queues.

One possible rationale for why queue length would affect capacity comes from the physics of traffic flow and, in particular, how changes in vehicle speed propagate

[6]Economists have long used instrumental variables in such settings to econometrically separate demand and supply (see, e.g., Angrist and Krueger, 2001). An alternative to this instrumental variables strategy would be to exclude observations affected by supply shocks. This ends up being impractical, however, because many supply shocks are not observed. For example, the California Department of Transportation's *Performance Measurement System (PeMS)* tracks major incidents but many smaller incidents are unreported.

through a queue of vehicles. Since Lighthill and Whitham (1955), transportation engineers have used "fluid dynamics" and the "method of kinematic waves" to model the movement of vehicles in a queue. The implication of these models is that longer queues are associated with lower average speeds, reducing overall capacity relative to models without the "theory of shock waves" (Mun, 1994, 1999). Our 2SLS regressions can be viewed as an attempt to measure empirically the quantitative importance of these physical mechanisms.

When estimating Equation (1) we restrict the sample to include only observations from periods in which there is a queue. By definition, if there is no queue, then traffic flows have not reached 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 with queue length as the independent variable we compare traffic flows between periods with different lengths of non-zero queues, and we refer to traffic flows during these periods as "capacity".

2.2.2 Event Study Analyses

Our second test takes the form of a standard event study regression. The event of interest in our context is the moment in time that the queue forms. These event study analyses allow us to assess whether there is a sudden drop in highway capacity

at the onset of a queue. In particular, we estimate regressions of the form:

$$\text{traffic flow}_t = \sum_{k=-16}^{16} \beta_k 1[\tau_t = k]_t + \omega_t. \tag{2}$$

The dependent variable in these regressions is traffic flow in 5-minute period t, measured downstream of the bottleneck. The independent variables of interest are a vector of event-time indicator variables. In particular, we construct a variable τ_t defined such that $\tau = 0$ for the exact moment in which the queue forms, $\tau = -16$ for 16 periods (i.e 80 minutes) before the queue forms, $\tau = 16$ for 16 periods (i.e. 80 minutes) after the queue forms, and so on. Our estimates of β_k summarize how traffic flows vary before and after the queue forms. We include no additional control variables, so although we estimate the regression using least squares, it is equivalent to taking conditional averages in event time.

This event study approach closely coincides with other models of hypercongestion that have appeared in the literature. Many studies find evidence of "capacity drop", "flow breakdown", or the "two capacity phenomenon" at bottlenecks.[7] This decrease in capacity could be caused, for example, by vehicles stopping prior to the merge or failing to efficiently fill gaps in the queue. Many studies take this capacity drop as given. For example, Hall (2018b) assumes that queues reduce capacity at bottlenecks by 10%.

[7]See, e.g., Banks (1990, 1991); Hall and Agyemang-Duah (1991); Persaud et al. (1998); Cassidy and Bertini (1999); Bertini and Malik (2004); Zhang and Levinson (2004); Chung et al. (2007); Oh and Yeo (2012). Hall (2018b) reviews this literature carefully, reporting that 16 out of 17 papers find evidence of a capacity drop, with a median capacity drop of 10%.

Unlike the 2SLS regressions, we do not restrict the sample to include only observations in which there is a queue, as that would omit observations before $\tau = 0$. We thus refer to the dependent variable in these regressions as traffic flow, rather than capacity. As we show later, however, the estimates of β_k provide some information about the degree of demand starvation. In our empirical applications we tend not to see large increases in flow leading up to queue formation, suggesting that flow is near capacity in the periods leading up to queue formation.

Our two empirical tests are complementary. Whereas the event study analyses focus on the transition between no queue and queue, the 2SLS regressions compare traffic flows with different lengths of non-zero queues. Thus the identifying variation used in the two tests is different, and in fact, completely disjoint. In the former, it is the mere formation of a queue, not the queue length, that matters for highway capacity, whereas in the latter, the reduction in capacity comes from the length of the queue. These are two alternative mechanisms for hypercongestion, but both imply that as demand for travel goes up, it causes a decrease in the capacity of the road system.

3 Empirical Application

3.1 Data

Our empirical analyses focus on three study sites. All three sites are in California, allowing us to use high-quality, comparable data from a single source, the Califor-

nia Department of Transportation (Caltrans). In particular our data come from Caltrans' statewide network of "loop detectors", which record information on both traffic flows and average vehicle speed.[8]

We selected these three sites based on several criteria. Most importantly, we wanted sites with a single, clearly identified bottleneck. In all three of our study sites there is a specific location where traffic slows and the queue forms, followed by a downstream location where traffic generally returns to full speed. We did not want sites with multiple bottlenecks, as it becomes difficult to assess the impact of any individual bottleneck. In addition, we wanted sites with good data coverage. Many promising sites were discarded because there was an insufficient number of well-functioning loop detectors nearby.

3.1.1 Westbound SR-24

Our first study site is the westbound direction of California State Route 24 (SR-24) at the Caldecott Tunnel. SR-24 connects suburban Contra Costa County, to the east, with the cities of Oakland and San Francisco, to the west. This site is a classic bottleneck, with the number of lanes decreasing as traffic approaches the tunnel. This is a site where traffic delays are common; indeed, transportation engineers have repeatedly studied this exact site (Chin and May, 1991; Chung and Cassidy, 2002;

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

Chung et al., 2007). 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 merged from four lanes to two as they approached the tunnel.[9]

Figure 2 shows the study site. 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 site and the location where the vehicle queue typically begins. The figure also indicates using small circles the locations of loop detectors. We observe a set of two loop detectors after the merge but before the tunnel, as well as a series of loop detectors upstream of the merge.[10] For westbound travelers there is no reasonable alternative to traversing the tunnel.[11]

[9]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 this period, the middle bore operated westward during morning hours, as commuters drove toward Oakland and San Francisco, and eastward during afternoon and evening hours, as commuters drove toward suburban Contra Costa County. Afternoon westbound traffic is lighter than eastbound traffic, but with only a single bore open in the westbound direction, the bottleneck was more than sufficient to generate significant traffic delays on weekday afternoons. We do not use the eastbound morning bottleneck in our analysis because it features traffic merging from multiple directions, making it difficult to measure queue length.

[10]For the event study analysis, the first upstream detector is approximately 1,000 feet from the bottleneck. This spacing introduces some delay between the formation of the queue and its detection. Detectors at other sites — in particular at I-15 — are located closer to their respective bottlenecks. Reassuringly, the event study analysis generates qualitatively similar findings across all three sites.

[11]For visual clarity the figure does not include exits and entrances. One of the significant advantages of this study site is that there are relatively few exits and entrances nearby. The last highway entrance prior to the bottleneck is approximately 9,000 feet (1.7 miles) east of the tunnel; the entrance at Gateway Blvd did not connect to any through roads. Subsequent to our sample dates, the Gateway Blvd exit was renamed Wilder Rd.

3.1.2 Southbound I-15

Our second study site is the southbound direction of Interstate 15 (I-15) northeast of San Diego. I-15 connects suburban San Diego County, to the north, with the city of San Diego and I-5, to the south. We focus on afternoon hours at the location where I-15 crosses I-805, another major north-south highway. As Figure 2 illustrates, I-15 southbound has five lanes prior to crossing I-805. However, while crossing I-805, I-15 reduces to only two lanes, before widening to three lanes. As we show, this bottleneck results in frequent queuing during afternoon hours. We focus in particular on afternoon hours between 2015 and 2018, years during which the relevant loop detectors were online and functioning reliably.

Of our three study sites, I-15 is the most complicated. As the figure suggests, there are significant flows both to and from I-805. For visual clarity the figure does not include all entrances and exits, but there are also entrances and exits at Adams Avenue, El Cajon Boulevard, and University Avenue. We examined loop detector data from these entrances and exits, as well as changes in net flows on I-15, and found that these entrances and exits involve flows that are small compared to the flows coming on and off of I-805. Nevertheless, it is important to corroborate results from I-15 with results from the other two sites where there is much less scope for substitution to alternative routes.[12]

[12]One advantage of the I-15 site is that the first upstream detector, At I-805, is located only 300 feet from the bottleneck. In the context of the event study analysis this means that any queue is detected almost immediately, since even emergency braking from freeway speeds requires up to 200 feet to stop.

3.1.3 Eastbound SR-12

Our third study site is the eastbound direction of California State Route 12 (SR-12). SR-12 runs through Sonoma, Napa, and Solano Counties, before merging with Interstate 80 (I-80), at which point drivers continue north toward Sacramento. We focus on afternoon hours at a location just west of I-80. As Figure 2 illustrates, at this location SR-12 merges from two lanes down to one lane.[13] As we show later, this merge results in queues that are often very long. This site is a classic bottleneck with no reasonable alternatives for eastbound drivers. We focus on 2017 and 2018, years during which the relevant loop detectors were online and functioning reliably.

3.2 Vehicle Flows

Figure 3 plots vehicle flows by hour-of-day for our three study sites. Each data series describes a different loop detector location. The legend orders detectors in the direction of traffic flow such that for each site, the last detector in the list corresponds to the farthest downstream detector (past the bottleneck). The unit of observation in the underlying data is a five-minute period. Throughout the analysis we average across lanes at a given detector location. In general, traffic flows and speeds tend to be highly correlated across lanes, as drivers arbitrage any differences.

Morning and afternoon commuting patterns are visible for all three sites. Total vehicle traffic peaks in the morning at SR-24, but as noted earlier we focus on afternoons

[13]The first upstream detector, W of Red Top Rd, is located approximately 700 feet from the bottleneck. This spacing is closer than on SR-24 but further than on I-15.

when the middle bore of the Caldecott Tunnel was operated in the opposite direction. In the afternoons vehicles merge from four lanes to two as they approach the tunnel, resulting in mean vehicle flows per lane that are approximately twice as high at the downstream location (Fish Ranch Rd) as compared to upstream locations.

At I-15 and SR-12, total vehicle traffic peaks in the afternoon. As with SR-24, the downstream detectors (S of I-805 and Red Top Rd respectively) register higher flows per lane as traffic enters the "neck" of the bottle. With SR-12, the downstream flows per lane (at Red Top Rd) are approximately twice as high as flows at the upstream location, reflecting the merge from two lanes to one lane.

3.3 Vehicle Speeds

Figure 4 plots vehicle speeds by hour-of-day for our three study sites. During afternoon hours there are dramatic decreases in average speeds at all three sites. Speeds tend to decrease the most at detectors just upstream of the bottleneck. For example, on SR-24 the detector immediately upstream of the bottleneck (Gateway Blvd) exhibits average speeds below 40 miles-per-hour between about 3pm and 6pm. With I-15 all six detectors experience large decreases in speed during afternoon hours. Finally, SR-12 has the most severe afternoon decreases in speed, with several upstream detectors exhibiting average speeds below 30, and even below 20 miles-per-hour.

Speeds tend to decrease much less at downstream detectors. On SR-24, for example, average speeds immediately upstream (Gateway) and downstream (Fish Ranch) track

each other closely throughout most of the day. Between 3pm and 6pm, however, there is a significant divergence; upstream speeds slow to below 20 miles-per-hour, while downstream speeds remain above 40 miles-per-hour. Similarly on SR-12, the upstream locations (W of Red Top and E Miners) slow down to below 20 miles-per-hour, while the downstream location (Red Top Rd) maintains average speeds above 40 miles-per-hour.

3.4 Speed-Flow Curves

Figure 5 plots vehicle speeds against traffic flows for our three study sites. We constructed these scatterplots using all five-minute observations from the immediate upstream detector from each site. For the SR-24 site, we plot data from 1pm to 11:55pm, hours during which the tunnel was operated eastbound; for all other sites we plot all available data. With several years of data for each site, each scatterplot includes many observations, so we use colors to reflect the density of observations in each cell.

The basic pattern is similar to the speed-flow curves in Figure 1. With all three sites there is a large mass of observations at 60 miles-per-hour or faster. Speeds decrease modestly with flow rates along the top part of the speed-flow curve. But then, as with Figure 1, there are also large numbers of low-speed, low-flow observations which make the curve appear to bend backward.

These low-speed, low-flow observations tend to occur during afternoon hours when there are vehicle queues. Low-speed, low-flow observations may also reflect transient

supply shocks, like road construction, lane closures, and stalled vehicles. Speeds and traffic flows are determined in equilibrium by interactions between demand and supply. Thus, in general, it does not make sense to interpret this locus of observations as a supply curve.

3.5 Queue Lengths

We now turn to focus more explicitly on vehicle queues. These patterns of vehicle speeds and traffic flows imply that there is significant queuing of vehicles occurring during afternoon hours. With a mild assumption we can use our data to measure the presence of vehicle queues more directly. In particular, we assume that a queue is present whenever traffic is moving at under 30 miles-per-hour. This threshold is arbitrary, but we show later that our results are robust to alternative definitions. This assumption allows us to both detect when a queue forms each day and to measure the length of the queue.

An example is helpful. For SR-24, the first upstream 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. We define a queue as being present if the average speed at Gateway Blvd is below 30 miles-per-hour. We then measure the length of the queue using the number of consecutive upstream detectors for which we observe speeds below 30 miles-per-hour.[14]

[14]For example, if the measured speed is below 30 miles-per-hour at the first upstream detector (Gateway), but not at the second (Orinda West), then we conclude that the queue is 1,690 feet in

Figure 6 plots mean queue lengths by hour-of-day. These are raw means by hour-of-day, with queue length coded as zero for periods without queues. During morning hours, there are almost never queues. During afternoon hours, however, long queues tend to form at all three study sites. Queues reach maximum length at about 6pm for all three sites. Lengths vary across sites, but queues can be very long; at SR-12, for example, the average queue between about 5pm and 6pm exceeds one mile.[15]

3.6 Estimation Sample

In the instrumental variables analyses we restrict the sample to include only periods in which a queue has formed. By definition, if there is no queue, then traffic flows have not reached maximum capacity, and measurements are limited by insufficient demand rather than maximum capacity.

Conditioning the estimation sample on the presence of a queue is an effective approach for addressing demand starvation. However, this sample selection criterion can introduce subtle but meaningful bias into our analysis. 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? One possibility is a negative supply shock, such as poor weather or an accident. When conditioning on a queue existing, negative supply shocks will thus be more common during periods

length. "Broken" queues (e.g. a case in which traffic moves below 30 miles-per-hour at Gateway, above 30 miles-per-hour at Orinda West, and below 30 miles-per-hour at Camino Pablo West) are rare in our estimation sample, and our results are robust to their inclusion or exclusion.

[15]The online appendix includes additional descriptive statistics about queues. In particular, Appendix Figure A2 presents histograms of queue length by site by year, and Appendix Figure A5 presents histograms of the time-of-day at which the queue begins each day.

of low demand than during periods of high demand, as supply shocks are necessary for the formation of queues during low demand periods. Thus, the sample selection criterion itself can generate a spurious positive relationship between capacity and queue length even if demand is randomly assigned.

To mitigate this bias we focus on periods when demand tends to be sufficient to generate a queue even in the absence of a negative supply shock. For all three sites the frequency of queuing peaks on non-holiday weekdays between 4pm and 7pm (see Appendix Figure A3). Accordingly, in our baseline estimation sample we restrict observations to these weekday hours.[16] We also report results in which we further narrow the focus to 4:30pm to 6:30pm, when the frequency of queuing is even higher.

For each study site we also report results restricting the sample to include only between 4:30pm and 6:30pm during summer months. The only supply shock evident to us that might correlate with time-of-day is lighting conditions. However, during summer months, which we define as June to August, the sunset generally occurs after 8pm, well after the 4:30pm and 6:30pm window. Moreover, traffic is much less likely during summer months to be impacted by adverse weather.[17]

There is also a related issue that occurs at our I-15 site only. Downstream of the two-lane bottleneck, I-15 absorbs inflows from I-805 and then traverses an interchange

[16]The online appendix presents summary statistics. In particular, Appendix Tables A1 and A2 present summary statistics of the complete data set and the 4pm-7pm sample with a queue present, respectively.

[17]For example, at the SR-24 study site, total monthly precipitation during summer months never exceeds 0.1 inches, and fog is rarely present on summer afternoons. At the SR-12 study site, sunset is never an issue, since drivers are going east.

with another highway, SR-94. On certain days queuing forms at these downstream locations, and the second bottleneck can back up to our study site, restricting traffic flow. To avoid potential problems, we exclude observations for which downstream average vehicle speeds (S of I-805) are below 35 miles-per-hour. At our other two sites, there is no downstream bottleneck, and we impose no such sample restriction.

Finally, before proceeding we note that our descriptive statistics provide an informal visual version of our instrumental variables test. Figure 6 revealed that queue lengths vary substantially between 4pm and 7pm. This is visual evidence of a strong "first-stage" relationship between queue length and time-of-day. In addition, Figure 3 revealed that vehicle flows downstream of the bottleneck are relatively constant between 4pm and 7pm, and Appendix Figure A4 confirms that this pattern holds after conditioning on a queue being present. The juxtaposition of Figures 3 and 6 provides evidence against hypercongestion. While queue length varies substantially between 4pm and 7pm, mean vehicle flows do not. The relatively flat time profile for traffic flows thus suggests there is no reduced-form relationship between time-of-day and highway capacity, implying that the instrumental variables estimate is likely to be close to zero.

4 Instrumental Variables

4.1 Baseline Estimates

Table 1 reports our baseline instrumental variables estimates for all three study sites. As we described earlier with Equation (1), the dependent variable in these regressions is capacity, measured downstream of the bottleneck in vehicles per five minutes. The independent variable of interest is queue length, measured in thousands of feet. 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 regressions are large in all cases, indicating that we do not have a weak instruments problem.

Across study sites and specifications, there is no evidence of hypercongestion. If hypercongestion were present, we would expect the estimates of α_1 to be negative. Instead, seven out of nine estimates are positive, and all are nine estimates are close to zero. For example, for SR-24 in Column (1), each additional 1,000 feet of queue is associated with a capacity increase of 3.6 vehicles per five minutes per lane. This is very small compared to the mean capacity per lane (176.2). When further restricting the sample in Columns (2) and (3), the estimates remain close to zero. The two negative estimates, both for SR-12, are tiny compared to mean capacity, in both cases less than 0.5%. Across sites and specifications, the estimates are sufficiently precise to rule out capacity drops of more than 2% per 1,000 feet of queue, even at 99% confidence levels.

4.2 Additional Specifications

Results are very similar across alternative specifications and robustness checks. When we estimate the queue length regressions using ordinary least squares, rather than 2SLS, the estimates tend to be negative, but they are extremely close to zero (Appendix Table A3). This pattern is consistent with queue length being driven by both demand and supply shocks — supply shocks introduce a spurious negative correlation between queue length and capacity, which is filtered out in our instrumental variables analyses. Regardless, the OLS estimates are precise enough to rule out capacity drops of more than 1% per 1,000 feet of queue across all sites and specifications.

Our results are also robust to using alternative thresholds to define queues. Recall that in our baseline specification in Table 1 we assume that a queue is present whenever traffic moves at under 30 miles-per-hour. Appendix Tables A4 and A5 present results using 25 miles-per-hour and 35-miles-per-hour as the threshold for a queue, respectively. Results with these alternative speed thresholds are nearly identical to our baseline results. Across study sites and specifications estimates are always close to zero, and in all cases we can rule out capacity drops of more than 2.5% per 1,000 feet of queue.

Thus across all specifications there is no evidence of hypercongestion. To the contrary, longer queues tend to be associated in many cases with slightly *higher* capacity. This could be, for example, 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.

5 Event Study Analysis

5.1 Visual Evidence

Figure 7 plots the results of our event study analyses. The horizontal axis is time in minutes relative to the onset of the queue, normalized so that the longest-duration afternoon queue begins at time zero on each day. As with the instrumental variable analyses, we define a queue as being present when upstream traffic is moving at under 30 miles-per-hour. For each weekday in our data, we identify the longest continuous period of queuing, and then take the first five-minute interval within that period to mark the onset of the queue. To focus on afternoons we exclude queues that do not start between 2:15pm and 7:00pm.

The event study analyses reveal no evidence of a decrease in capacity. For all three sites, capacity is essentially flat throughout, with no discontinuous change at the moment the queue is formed. Figure 7 also includes 95% confidence intervals, and these intervals are narrow enough to rule out even modest changes in capacity. To illustrate this, we included a simulated 10% capacity drop at queue onset in each panel. The 10% drop was chosen arbitrarily, but it is well within the range of estimates in the existing literature. The discordance between the two series indicates that we can rule out a capacity drop of this magnitude, or even considerably smaller magnitude.

5.2 Estimates and Standard Errors

Table 2 reports corresponding estimates and standard errors. As with Figure 7, these estimates are based on three separate event study regressions, one for each site. In Column (1) we report the change in capacity between the five minutes prior to queue formation and the five minutes after queue formation. That is, we calculate the difference between the last estimated β before queue formation (β_{-1}) and the first estimated β after queue formation (β_0). Columns (2) and (3) then expand the comparison to consider 20 and 30 minute symmetric windows, respectively. In these cases we calculate the difference in the average estimated β coefficients before and after queue formation, in order to report the implied change in capacity per five minutes.

Across study sites and specifications the estimates are very close to zero. Consistent with the visual evidence in Figure 7, Table 2 reveals no evidence of a decrease in capacity when the queue forms. Positive estimates indicate an increase in capacity. For example, for SR-24 in Column (1), we find that queue formation is associated with a capacity increase of 1.2 vehicles per five minutes. This is less than one percent of mean capacity. Results are similar with alternative windows and for the other sites — there is a mix of positive and negative estimates, but all are very small relative to mean capacity. These estimates are less precisely estimated than the instrumental variables estimates, but for all nine estimates in Table 2 we can rule out a 5% capacity drop or larger with 99% confidence. If we average the estimates in Column (1) across the three sites, we can reject a mean capacity drop across sites of 1% or larger.

This lack of evidence of a capacity drop stands in contrast to the vast majority of previous studies. For example, Zhang and Levinson (2004) reach a different conclusion examining bottlenecks in Minnesota. They find that the onset of the queue leads capacity to decrease by between 2% and 11% across sites. Like our study, they measure capacity downstream of the bottleneck, not within the queue. However, in their analysis they define the start of the queue as following an interval in which traffic flow exceeds its long-run average, and both upstream and downstream locations are uncongested. One concern with this approach is mean reversion, as average flows will tend to drop following an interval conditioned on having abnormally high flows.

Appendix Table A6 reports results from alternative event study analyses in which we estimate the specification used in Table 2 using median regressions. These estimates address the potential concern that our results are driven by large outliers, in either the positive or negative directions. Consistent with our baseline event study results, the median regression estimates are again close to zero, providing no evidence of a drop in capacity when the queue forms. Five of the nine estimates are positive, and in all cases we can reject a 5% capacity drop or larger.

6 Discussion

6.1 Effect Size Magnitudes

Neither the instrumental variables regressions nor the event study analyses show a decrease in highway capacity during periods of high demand. To put these results in context, Figure 8 plots a histogram of all of our estimates. We include all coefficient estimates from Tables 1 and 2, as well as from alternative analyses in Appendix Tables A3, A4, A5, and A6 (54 coefficients in total). These estimates summarize the results from both of our tests of hypercongestion across three sites and a rich variety of different specifications.

All of our estimates are clustered tightly around zero. To illustrate this fact we have included in Figure 8 two vertical lines. The righthand vertical line corresponds to the average observed capacity across all sites and tables — 154 vehicles per lane per five minutes. The lefthand vertical line corresponds to a hypothetical 10% decrease in capacity (15.4 vehicles per lane per five minutes). Even the most negative of our 54 estimates fall well short of this 10% threshold, and the vast majority of estimates are either positive or represent less than a 1% decrease in average capacity.

6.2 Policy Implications

We consider the policy implications of our results in the context of a rich existing literature that has examined the implications of hypercongestion using variations of

the "bottleneck" model. In this model drivers face a tradeoff between time delays and schedule inflexibility and optimize their departure times accordingly (Vickrey, 1969; Small, 1982; Arnott et al., 1990, 1993, 1994).

Economists have long recognized that traffic congestion represents a negative externality (Pigou, 1920; Vickrey, 1963, 1969; Newbery, 1990). When a motorist drives on a congested road, she decreases the average speed of all drivers, imposing an external cost. Our results imply, however, that this externality is not exacerbated by an additional decrease in capacity. Driving reduces average speeds, but we find no evidence that the travel supply curve is backward-bending. Thus our results imply that the marginal damages from driving are lower than would be implied by a supply curve exhibiting hypercongestion.

It is less clear what our results imply for optimal "Pigouvian" congestion pricing. Starting from an unregulated equilibrium, marginal damages are clearly lower without hypercongestion. However, at the social optimum there is less driving during peak times, so marginal damages are lower and typically queueing is avoided altogether (see, e.g. Mun, 1999; Fosgerau and Small, 2013). Thus whether or not hypercongestion exists likely has minimal impact on the how taxes are set in the optimal Pigouvian solution, as there may be no congestion at all.

This intuition is borne out in the existing literature. Arnott et al. (1993), for example, describes a model with a continuum of identical drivers facing a tradeoff between time delays and schedule inflexibility. In the optimal Pigouvian solution, drivers pay a time-varying tax that makes them indifferent between all departure times. This tax

depends on drivers' tastes for arriving early or late, but there is no queueing at the social optimum, so whether or not hypercongestion exists is irrelevant for setting the tax. With hypercongestion the welfare gains from optimal congestion pricing are larger, however, as total social costs are higher in the unregulated equilibrium.

Two recent papers by Jonathan Hall find that introducing driver heterogeneity does not change this basic intuition (Hall, 2018a,b). For example, Hall (2018a) structurally estimates drivers' preferences and then solves for optimal congestion pricing outcomes with different levels of hypercongestion. Counterfactual analyses (e.g. Table 5) show that gains from congestion pricing are larger when there is more hypercongestion, again because total social costs in the unregulated equilibrium increase with hypercongestion.

7 Conclusion

The concept of hypercongestion has influenced transportation economics models for over five decades. Our paper proposes two empirical tests of hypercongestion. Both tests are novel in the literature but correspond closely with how hypercongestion has been modeled in previous analytical studies. We apply both tests to high-quality data from three highway bottlenecks in California. Across tests, study sites, and specifications, we find no evidence of a reduction in highway capacity during periods of high demand.

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 equivalent. Speed is readily apparent to drivers, but capacity requires careful measurement.

On highways the feeling of being trapped in heavy traffic often occurs in a queue, waiting to pass a bottleneck. By definition the capacity *per lane* must drop when approaching a bottleneck as the number of lanes decreases. Nevertheless, we find that the capacity of the bottleneck itself — the rate at which vehicles pass through the bottleneck — does not drop when the queue first forms nor when the queue grows in length. No matter how much travel demand increases, the number of vehicles passing through the bottleneck remains stable.

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.

Allen, BL, FL Hall, and MA Gunter, "Another Look at Identifying Speed-Flow Relationships on Freeways," *Transportation Research Record*, 1985, *1005*, 54–64.

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, "A Bathtub Model of Downtown Traffic Congestion," *Journal of Urban Economics*, 2013, *76*, 110–121.

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

_ , _ , and _ , "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 1993, pp. 161–179.

_ , 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 Matthew A Turner, "Speed," *Review of Economics and Statistics*, 2018, *100* (4), 725–739.

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

Drake, JL and Joseph L Schofer, "A Statistical Analysis of Speed-Density Hypotheses," *Highway Research Record*, 1966, *154*, 53–87.

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

Fosgerau, Mogens and Kenneth A Small, "Hypercongestion in Downtown Metropolis," *Journal of Urban Economics*, 2013, *76*, 122–134.

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

Hall, Jonathan D., "Can Tolling Help Everyone? Estimating the Aggregate and Distributional Consequences of Congestion Pricing," *University of Toronto Working Paper*, 2018.

_ , "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways," *Journal of Public Economics*, 2018, *158*, 113–125.

Hanna, Rema, Gabriel Kreindler, and Benjamin A Olken, "Citywide Effects of High-Occupancy Vehicle Eestrictions: Evidence from "Three-in-One" in Jakarta," *Science*, 2017, *357* (6346), 89–93.

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

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.

Kreindler, Gabriel, "The Welfare Effect of Road Congestion Pricing: Experimental Evidence and Equilibrium Implications," *MIT Working paper*, 2018.

Lighthill, Michael James and Gerald Beresford Whitham, "On Kinematic Waves II. A Theory of Traffic Flow on Long Crowded Roads," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 1955, *229* (1178), 317–345.

Lindsey, C Robin and Erik T Verhoef, "Congestion Modeling," in "Handbook of Transportation Modelling" Elsevier Science 2008.

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

May, Adolf D, *Traffic Flow Fundamentals*, Prentice Hall, 1990.

Mun, Se-IL, "Traffic Jams and the Congestion Toll," *Transportation Research Part B: Methodological*, 1994, *28* (5), 365–375.

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

Newbery, David M, "Cost Recovery from Optimally Designed Roads," *Economica*, 1989, *56* (222), 165–185.

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

Ni, Daiheng, *Traffic Flow Theory: Characteristics, Experimental Methods, and Numerical Techniques*, Elsevier: Butterworth-Heinemann, 2015.

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.

_ , "The Bottleneck Model: An Assessment and Interpretation," *Economics of Transportation*, 2015, *4* (1-2), 110–117.

_ **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.

Sugiyama, Yuki, Minoru Fukui, Macoto Kikuchi, Katsuya Hasebe, Akihiro Nakayama, Katsuhiro Nishinari, Shin ichi Tadaki, and Satoshi Yukawa, "Traffic Jams Without Bottlenecks— Experimental Evidence for the Physical Mechanism of the Formation of a Jam," *New Journal of Physics*, 2008, *10* (3), 033001.

Tadaki, Shinichi, Macoto Kikuchi, Minoru Fukui, Akihiro Nakayama, Katsuhiro Nishinari, Akihiro Shibata, Yuki Sugiyama, Taturu Yosida, and Satoshi Yukawa, "Phase Transition in Traffic Jam Experiment on a Circuit," *New Journal of Physics*, 2013, *15* (10), 103034.

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

Transportation Review Board, "Highway Capacity Manual, Sixth Edition," *National Academies of Sciences, Engineering, Medicine: Washington, DC*, 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.

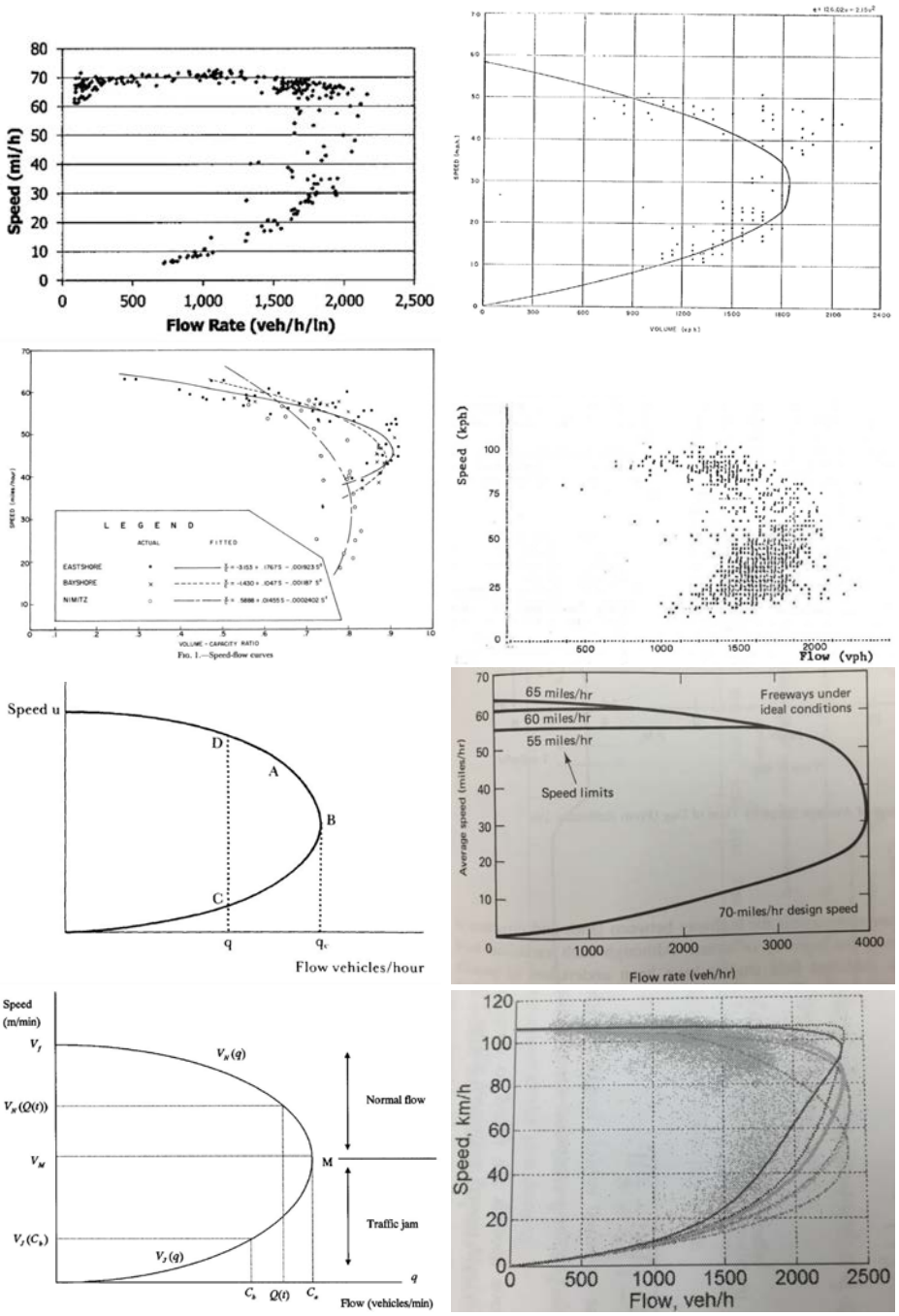
Walters, Alan A, "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, 1961, pp. 676–699.

Yang, Hai and Hai-Jun Huang, "Analysis of the Time-Varying Pricing of a Bottleneck with Elastic Demand Using Optimal Control Theory," *Transportation Research Part B: Methodological*, 1997, *31* (6), 425–440.

Yang, Jun, Avralt-Od Purevjav, and Shanjun Li, "The Marginal Cost of Traffic Congestion and Road Pricing: Evidence from a Natural Experiment in Beijing," *Working Paper*, 2018.

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

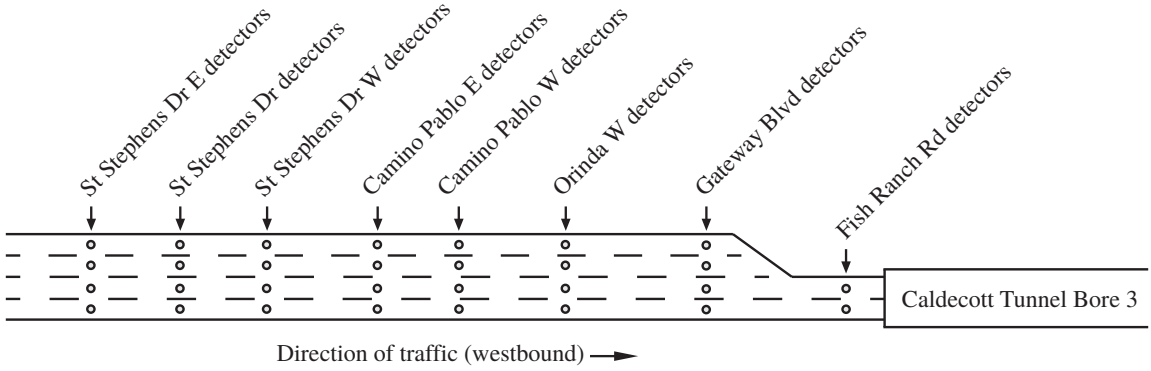
Figure 1: Speed-Flow Curves



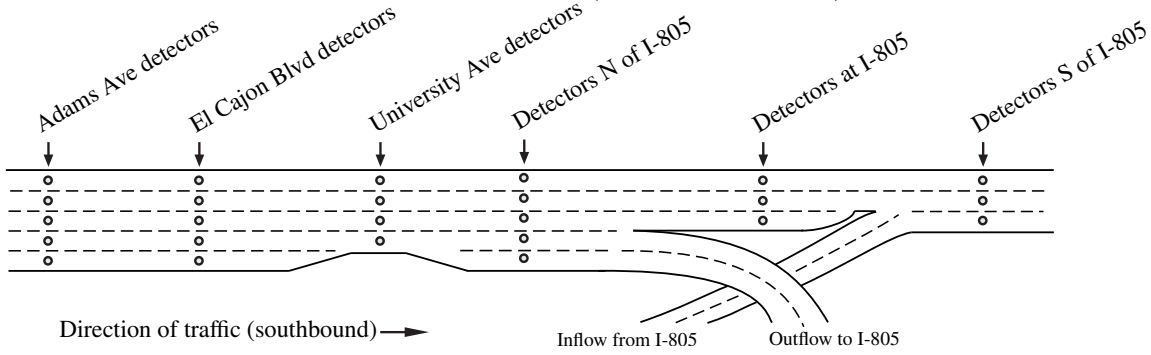
Sources: Transportation Research Board (2016), Drake and Schofer (1966), Keeler and Small (1977), Allen et al. (1985), Newbery (1989), May (1990), Mun (1999), and Ni (2015).

Figure 2: Study Sites

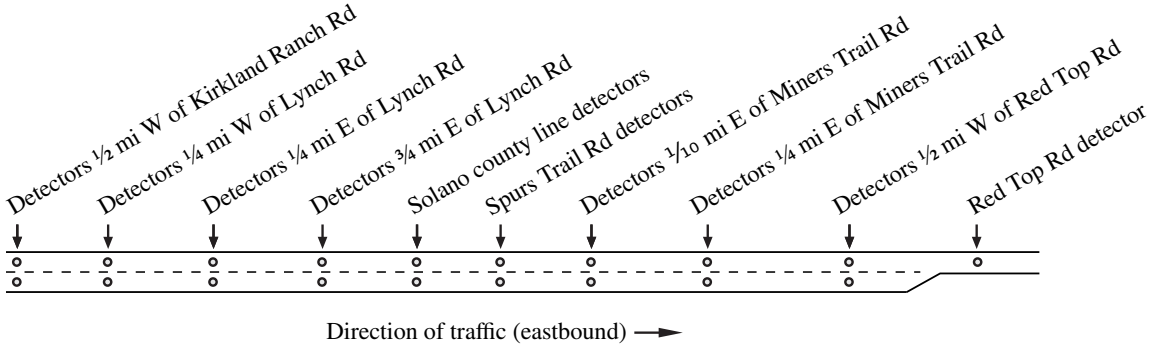
A. Westbound SR-24 (Caldecott Tunnel)



B. Southbound I-15 (Merge with I-805)

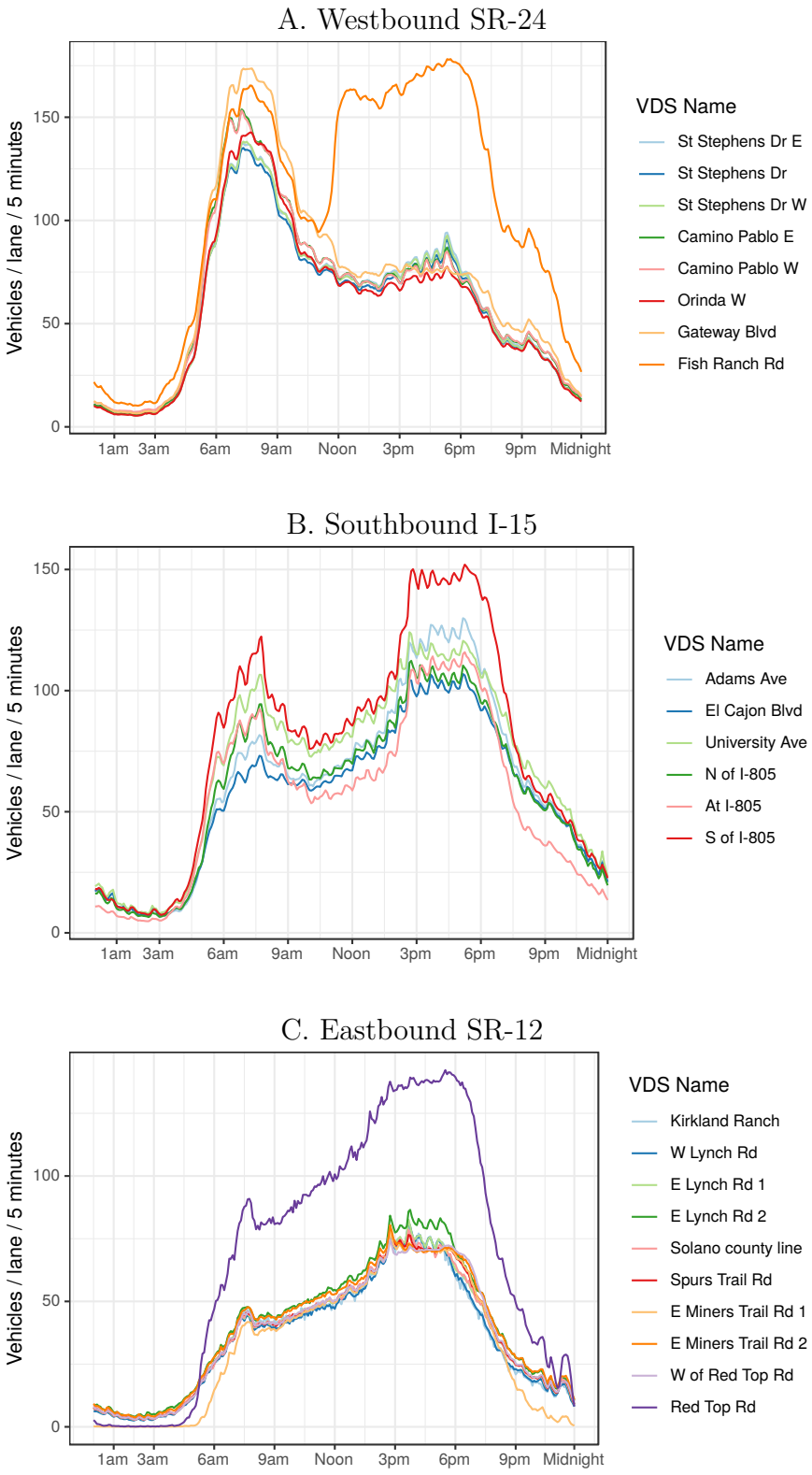


C. Eastbound SR-12 (Lane Reduction)



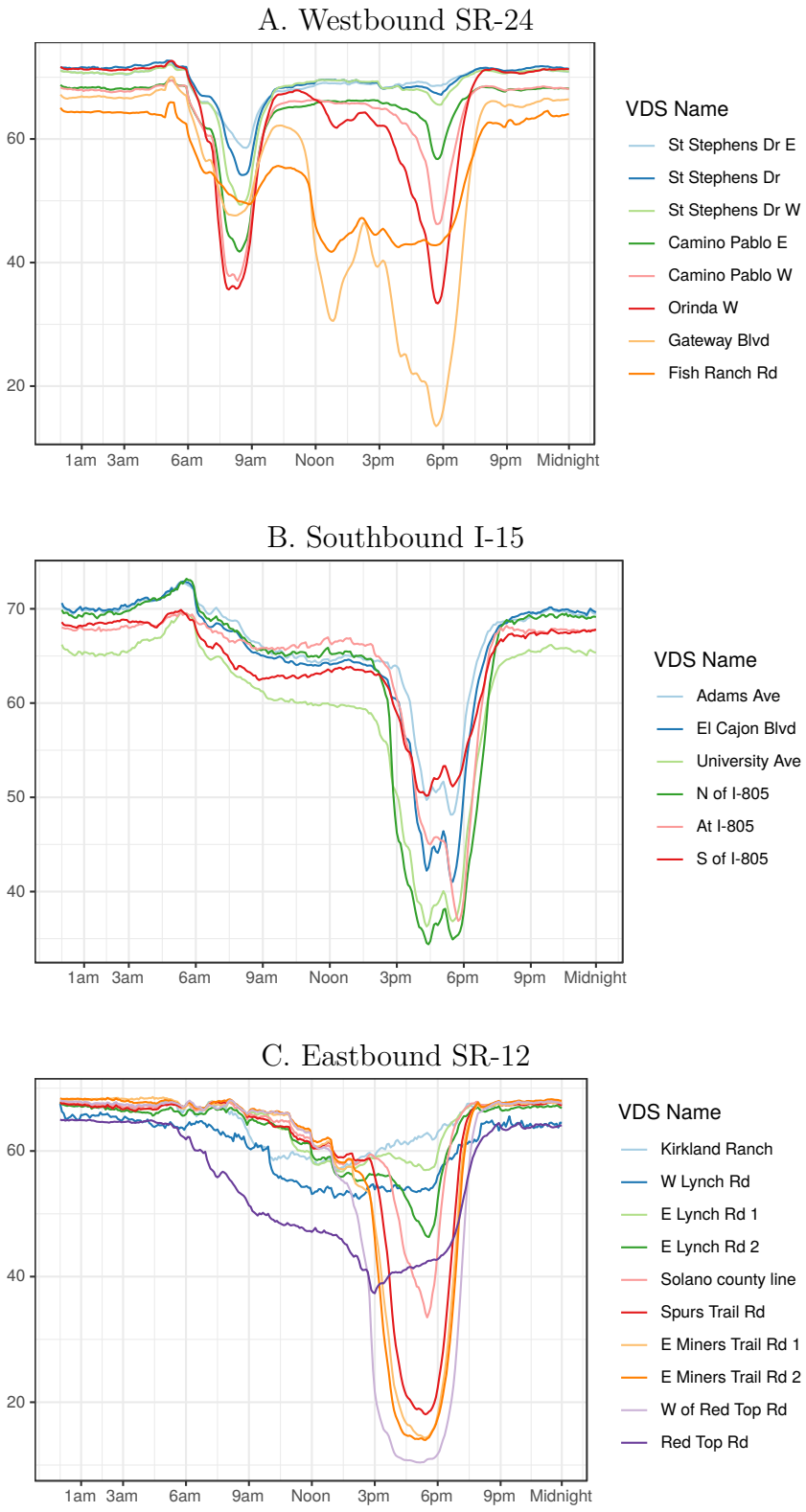
Notes: Figures approximately to scale.

Figure 3: Mean Traffic Flow



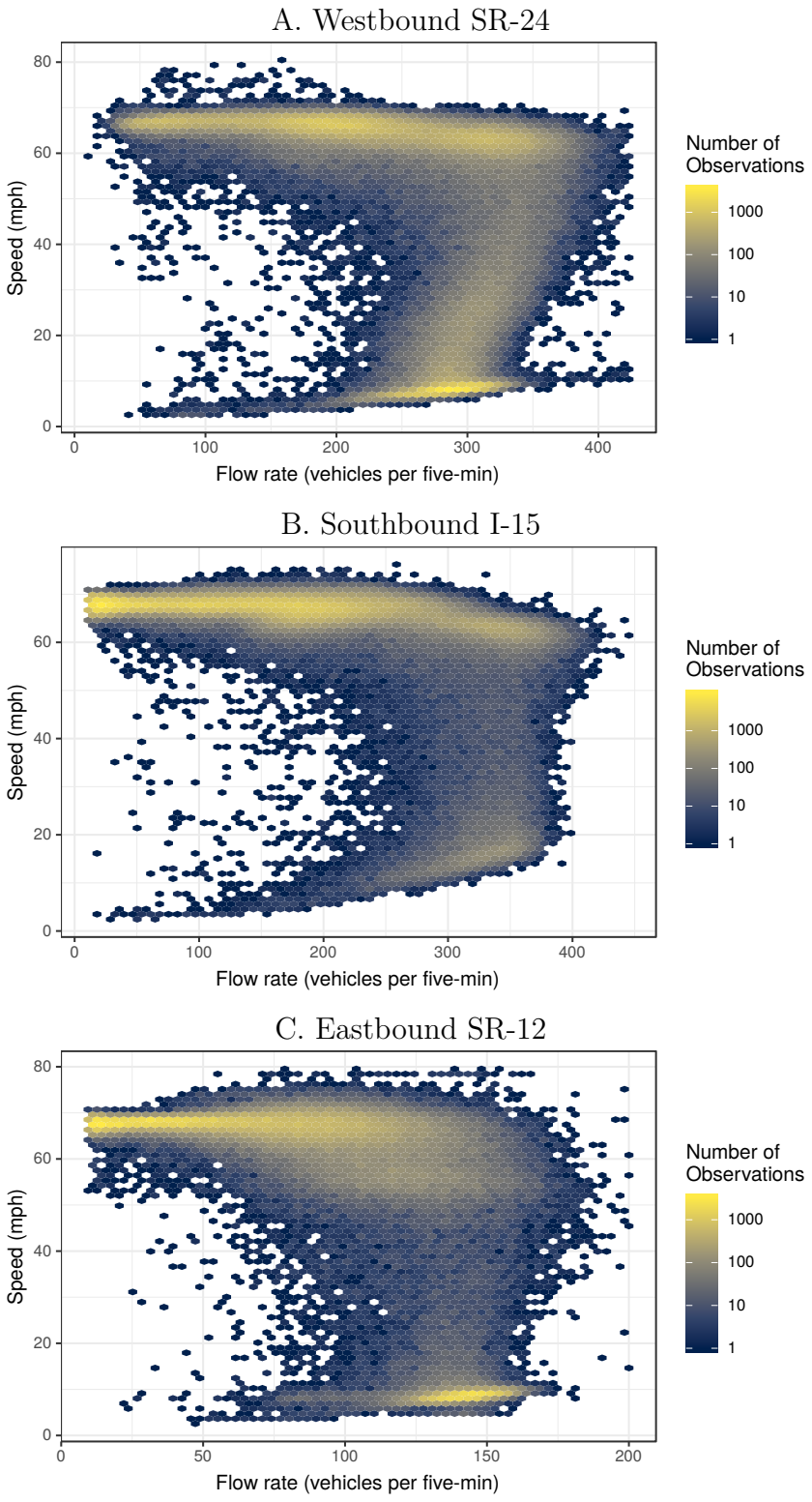
Notes: We exclude weekends and holidays.

Figure 4: Mean Vehicle Speed (in Miles-Per-Hour)



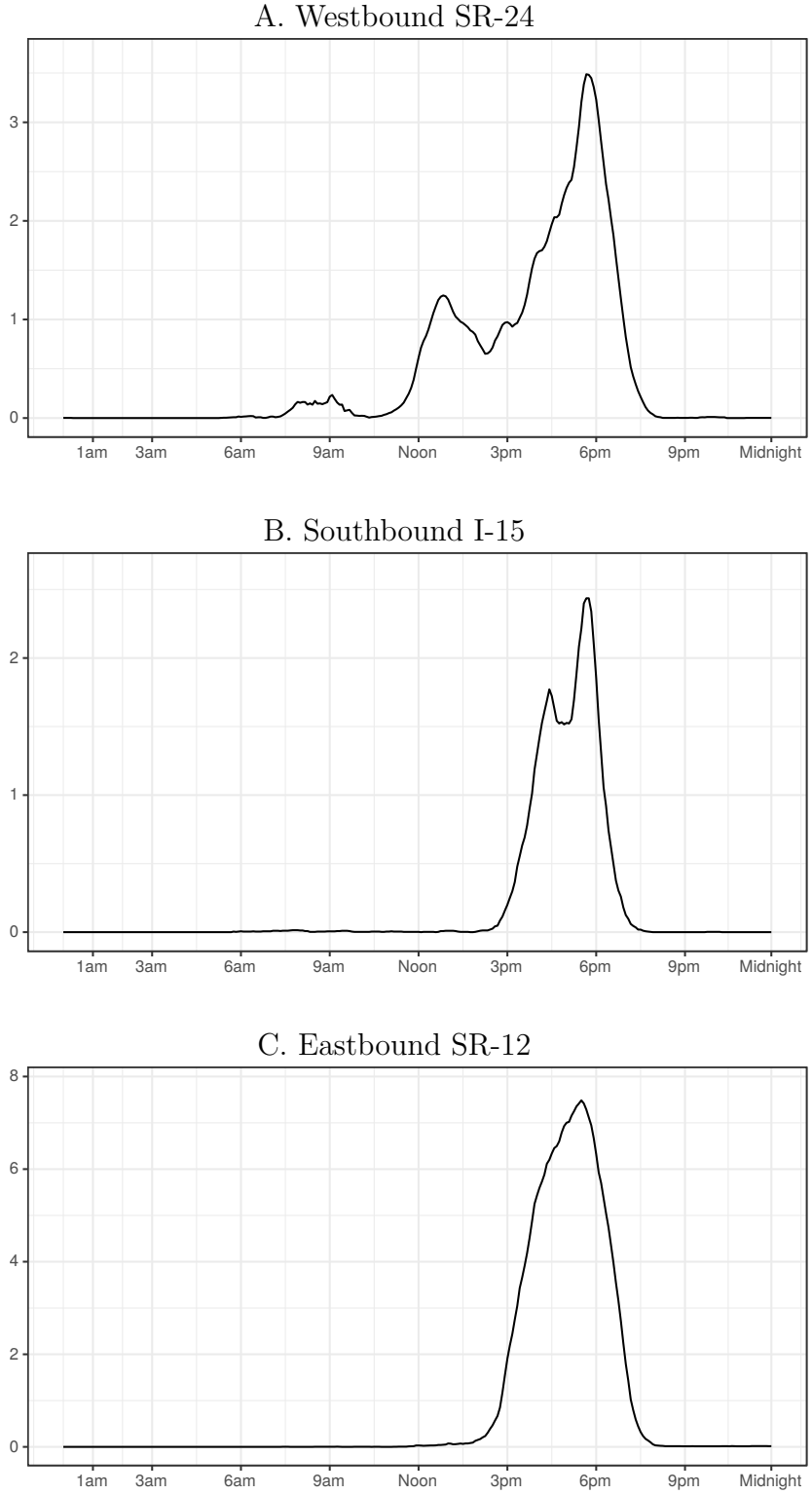
Notes: We exclude weekends and holidays.

Figure 5: Observed Speed–Flow Curves



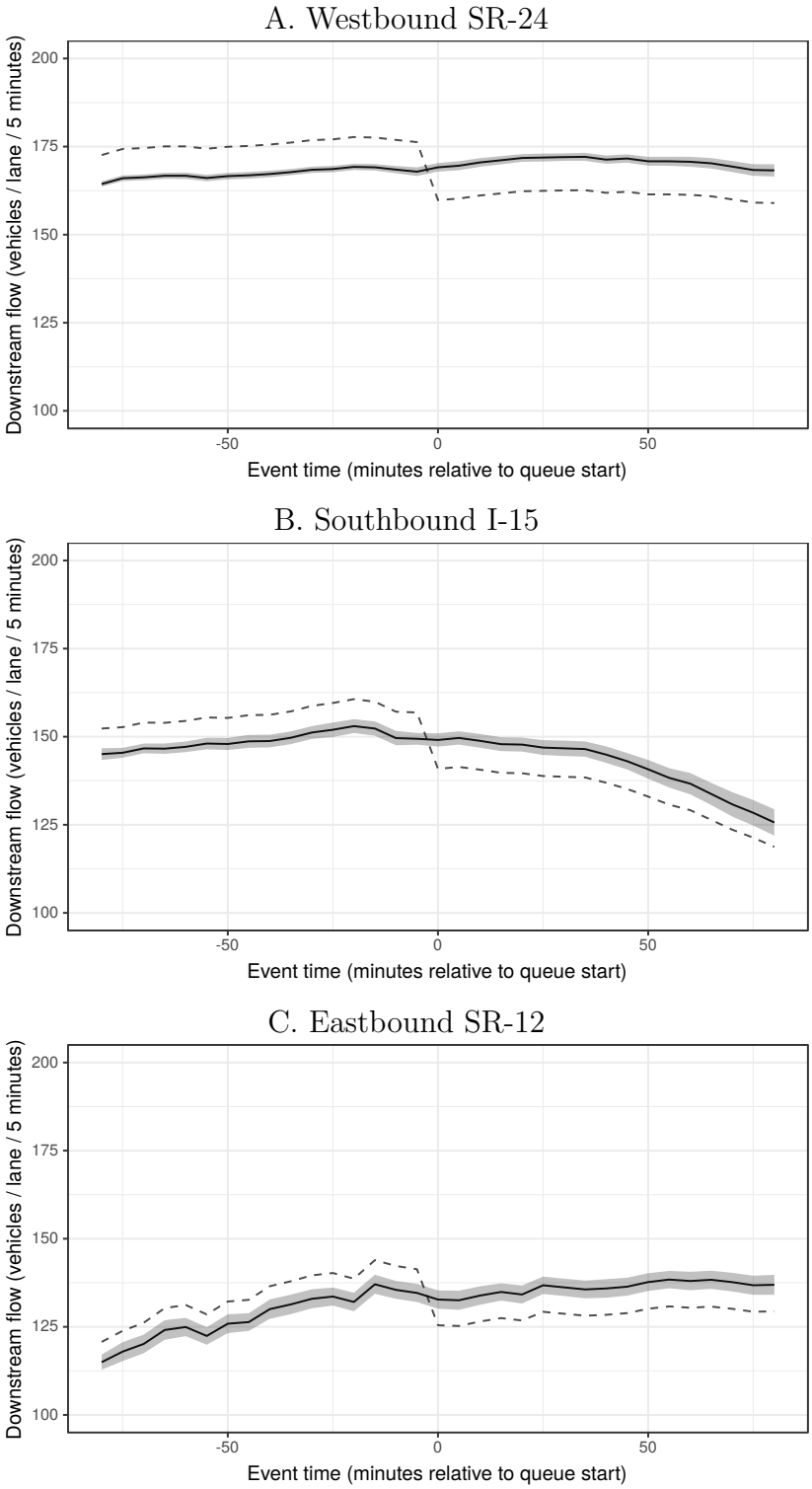
Notes: The x- and y-axes represent traffic flow and mean vehicle speeds, respectively, averaged across all lanes at the detector immediately upstream of each bottleneck. We exclude weekends and holidays. Colors represent the number of observations in each cell, as indicated in the legend.

Figure 6: Queue Length (in Thousands of Feet), Within Day



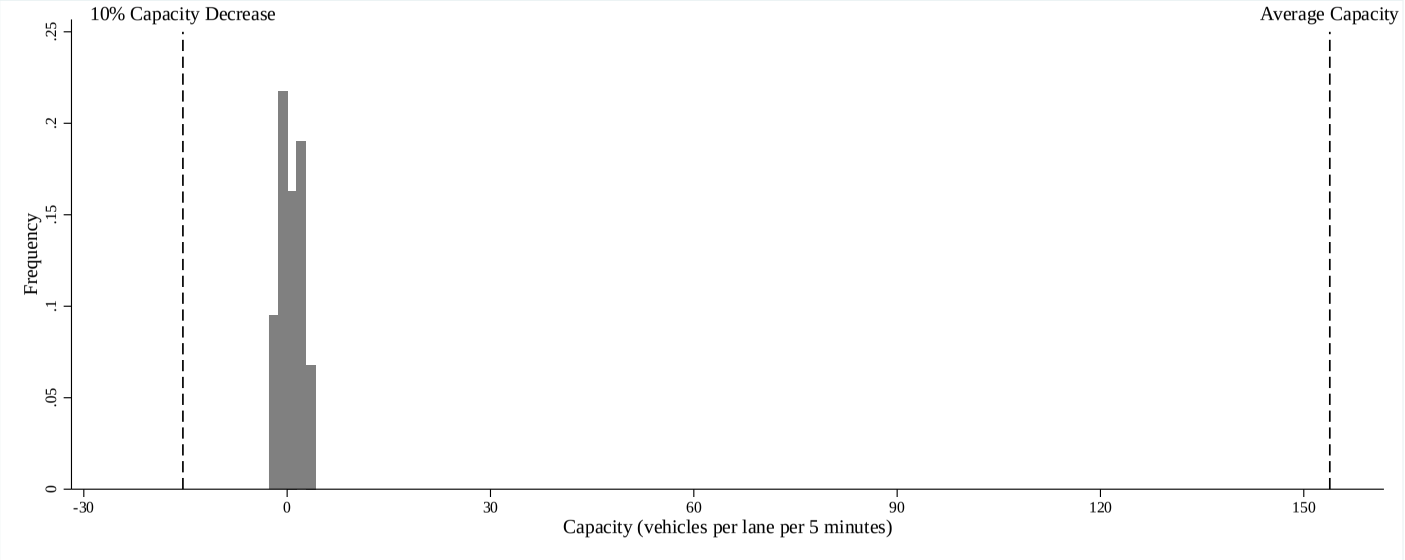
Notes: We plot for each study site the mean queue length (in thousands of feet) by 5-minute period. We exclude weekends and holidays. We define a queue as traffic moving under 30 miles-per-hour. Periods without queues are coded as having a queue of length zero.

Figure 7: Traffic Flows by Time of Queue Onset



Notes: These event study figures plot average vehicle flows in the 80 minutes before and after queue formation. Time is normalized so that the longest-duration afternoon queue begins at time zero on each day. The solid line plots mean capacity, with the shaded area representing a 95% confidence interval. The dashed line plots what mean capacity would look like if there were a capacity drop of 10% at queue onset, simulated by a drop from 5% above observed flows to 5.5% below observed flows at event time zero.

Figure 8: Distribution of Estimates



Notes: This figure plots the distribution of all 54 coefficient estimates from Tables 1, 2, A3, A4, A5, and A6. The righthand vertical line corresponds to the average observed capacity across all sites and aforementioned tables. The lefthand vertical line corresponds to a hypothetical 10% decrease in capacity.

Table 1: The Effect of Queue Length on Highway Capacity, 2SLS

	(1)	(2)	(3)
	Weekdays 4-7pm	Weekdays 4:30-6:30pm	Weekdays 4:30-6:30pm June-Aug
		A. Westbound SR-24	
Queue Length	3.581	2.017	2.557
	(0.365)	(0.403)	(0.545)
Total Observations	26,159	19,024	5,285
Number of Days	952	944	268
Dependent Variable Mean	176.2	177.1	177.8
Queue Length Mean	3.20	3.38	2.97
K-P F-stat	84.6	72.7	25.2
		B. Southbound I-15	
Queue Length	2.009	2.467	0.163
	(0.569)	(0.558)	(0.503)
Total Observations	7,718	6,414	1,616
Number of Days	687	660	160
Dependent Variable Mean	149.2	149.5	153.6
Queue Length Mean	3.96	3.99	3.67
K-P F-stat	30.5	25.4	7.31
		C. Eastbound SR-12	
Queue Length	0.102	-0.156	-0.681
	(0.188)	(0.275)	(0.573)
Total Observations	9,750	6,803	1,335
Number of Days	297	295	58
Dependent Variable Mean	139.2	139.6	139.8
Queue Length Mean	6.77	7.27	6.60
K-P F-stat	117.0	75.1	33.0

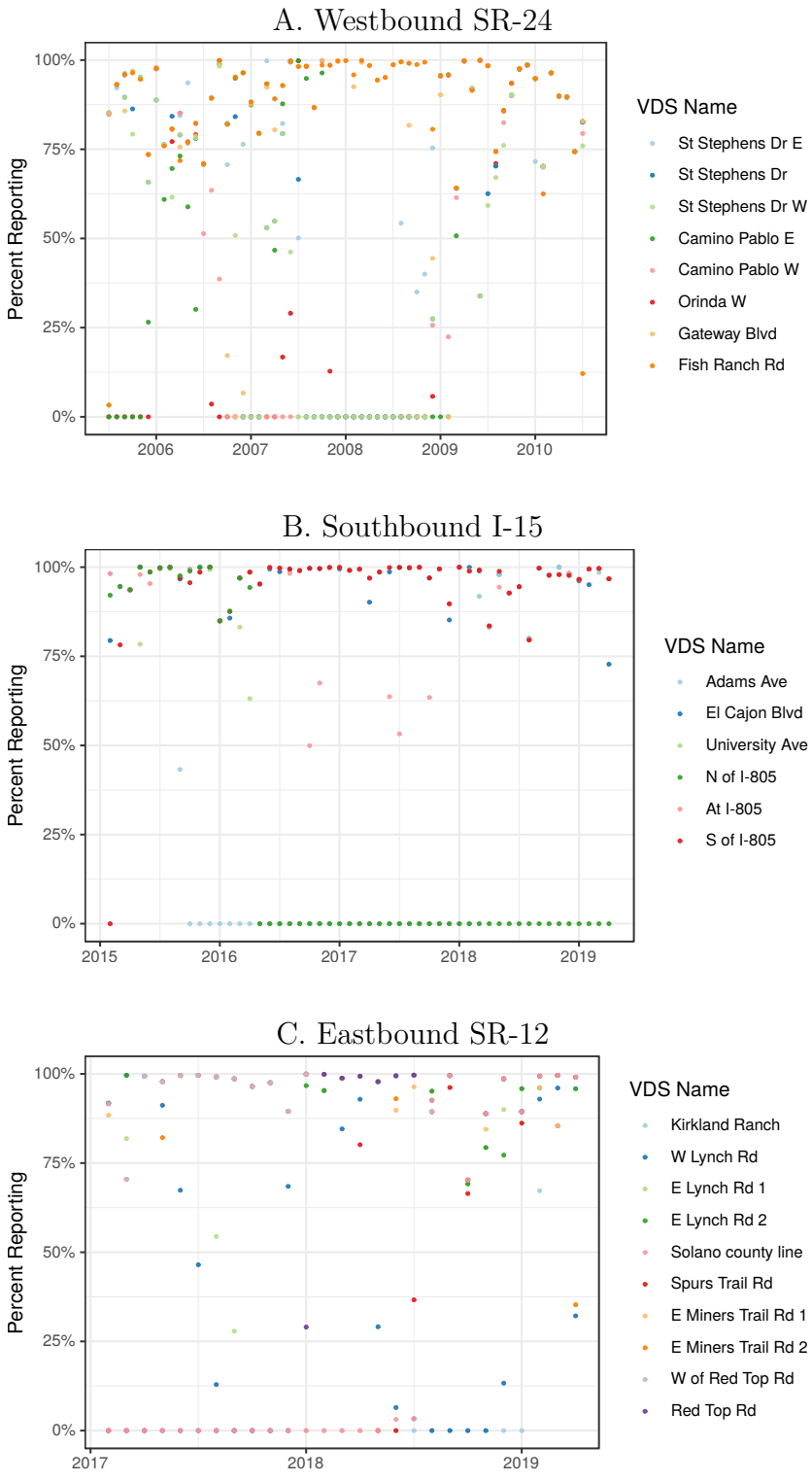
Notes: This table reports estimates and standard errors from nine separate regressions, all estimated using two-stage least squares (2SLS) with a third-order polynomial in time-of-day as the instruments. The dependent variable in all regressions is highway capacity, measured in vehicles per five minutes per lane, downstream of the bottleneck. Queue length is measured in thousands of feet. The sample includes all five-minute periods with a queue. Standard errors are clustered by day-of-sample.

Table 2: Estimated Change in Highway Capacity at Queue Formation

	(1) Ten Minute Window	(2) Twenty Minute Window	(3) Thirty Minute Window
		A. Westbound SR-24	
	1.232	1.167	1.240
	(0.357)	(0.353)	(0.366)
In-window mean	168.5	168.8	169.1
Number of days	706	706	706
		B. Southbound I-15	
	0.759	0.345	-1.680
	(0.433)	(0.404)	(0.386)
In-window mean	142.0	142.9	144.2
Number of days	694	694	694
		C. Eastbound SR-12	
	-1.879	-2.417	-2.679
	(1.090)	(0.865)	(0.720)
In-window mean	133.7	133.8	134.4
Number of days	247	247	247

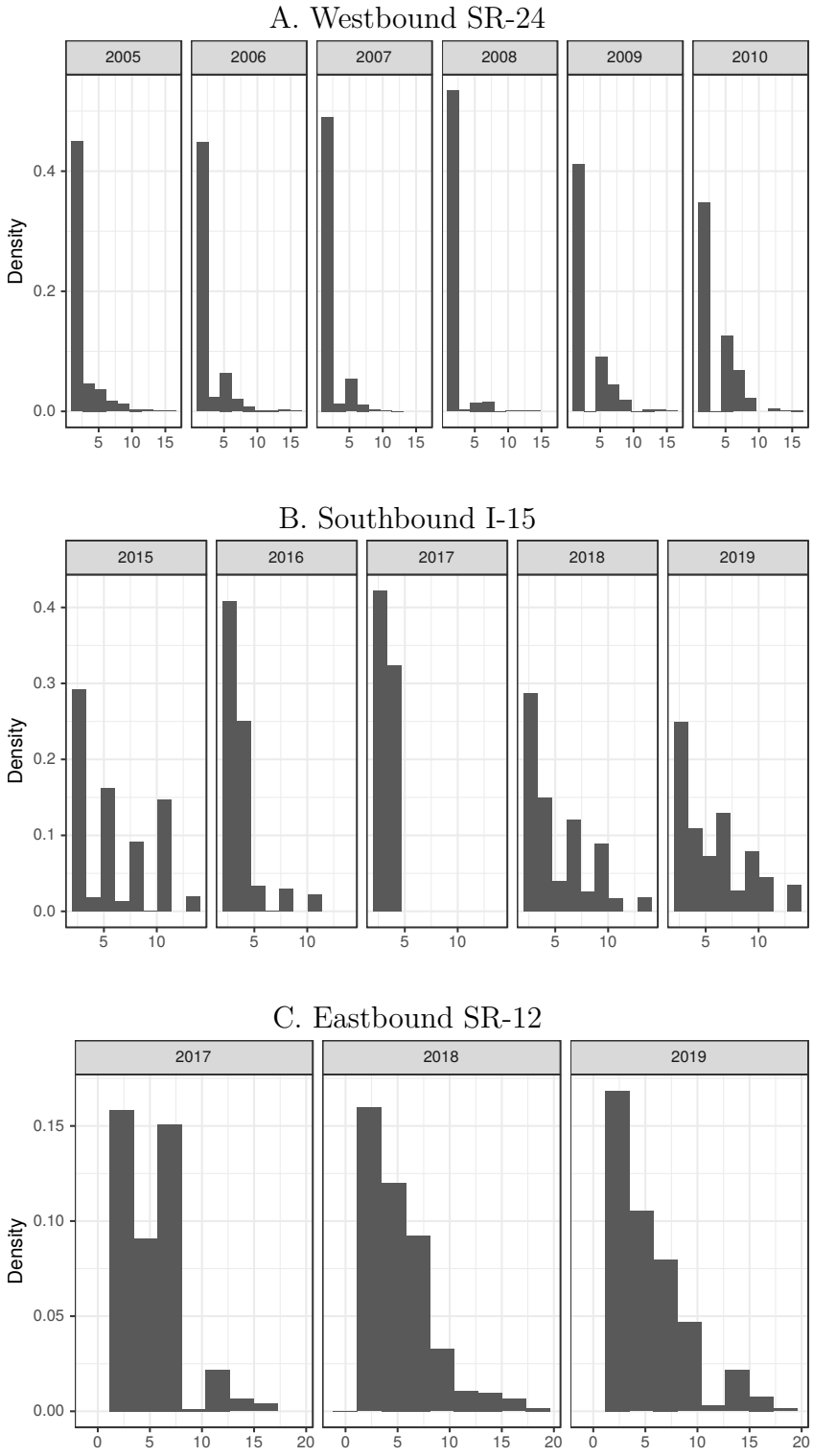
Notes: This table reports nine estimates of the change in highway capacity at queue formation. These estimates are based on coefficients from three separate event study regressions, one for each site. The dependent variable in all regressions is traffic flow (in vehicles per five minutes per lane), which we refer to as capacity because we focus on periods of queue formation when these bottlenecks operate at close to capacity. In Column (1) we report the change in capacity between the five minutes before queue formation and the five minutes after queue formation. Columns (2) and (3) expand the comparison to consider 20 and 30 minute symmetric windows (10 and 15 minutes in each direction), respectively. Standard errors are clustered by day-of-sample.

Figure A1: Detector Health by Month-of-Sample



Notes: These plots report detector health for detectors in our analysis. Each dot represents one month at one detector. The y-axis measures the fraction of intervals in the month, including weekends and holidays, for which the detector reported speed for all lanes. At each site, detector names are ordered from upstream to downstream.

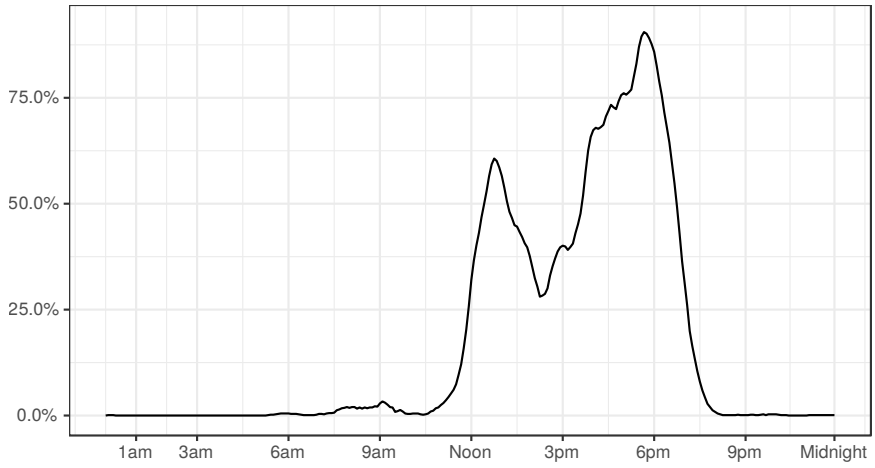
Figure A2: Queue Length (in Thousands of Feet), Histogram



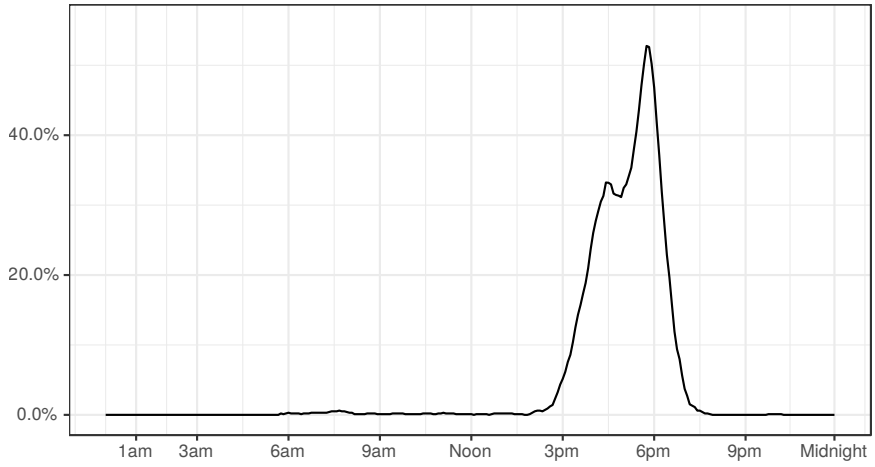
Notes: These histograms plot the distribution of queue lengths, for all observations when a queue is present, across the date ranges at each site (all available days from 1 June 2005 to 30 June 2010 for SR-24, 1 January 2015 to 31 March 2019 for I-15, and 1 January 2017 to 31 March 2019 for SR-12). Queue lengths are measured in thousands of feet. In each five-minute interval, queue length is calculated by totaling up the distances from each detector reporting speed less than 30 miles-per-hour to the next downstream detector. Detectors missing speed for more than one lane do not contribute to queue length.

Figure A3: Percentage of Hours With a Queue Present

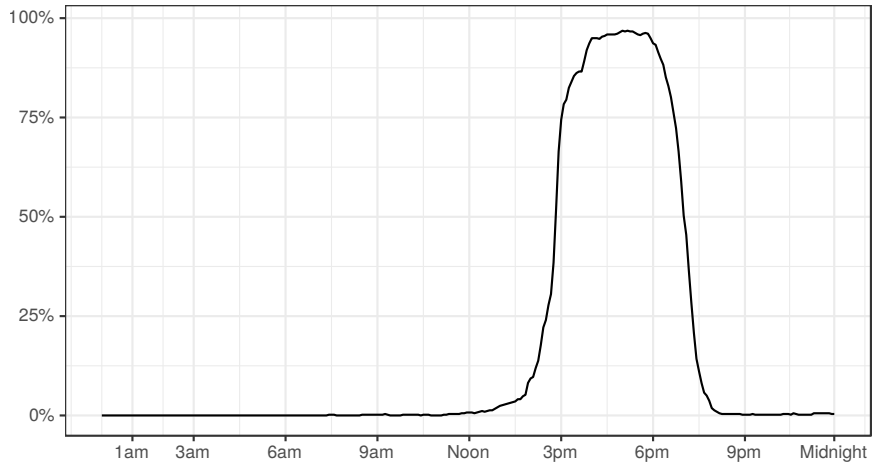
A. Westbound SR-24



B. Southbound I-15

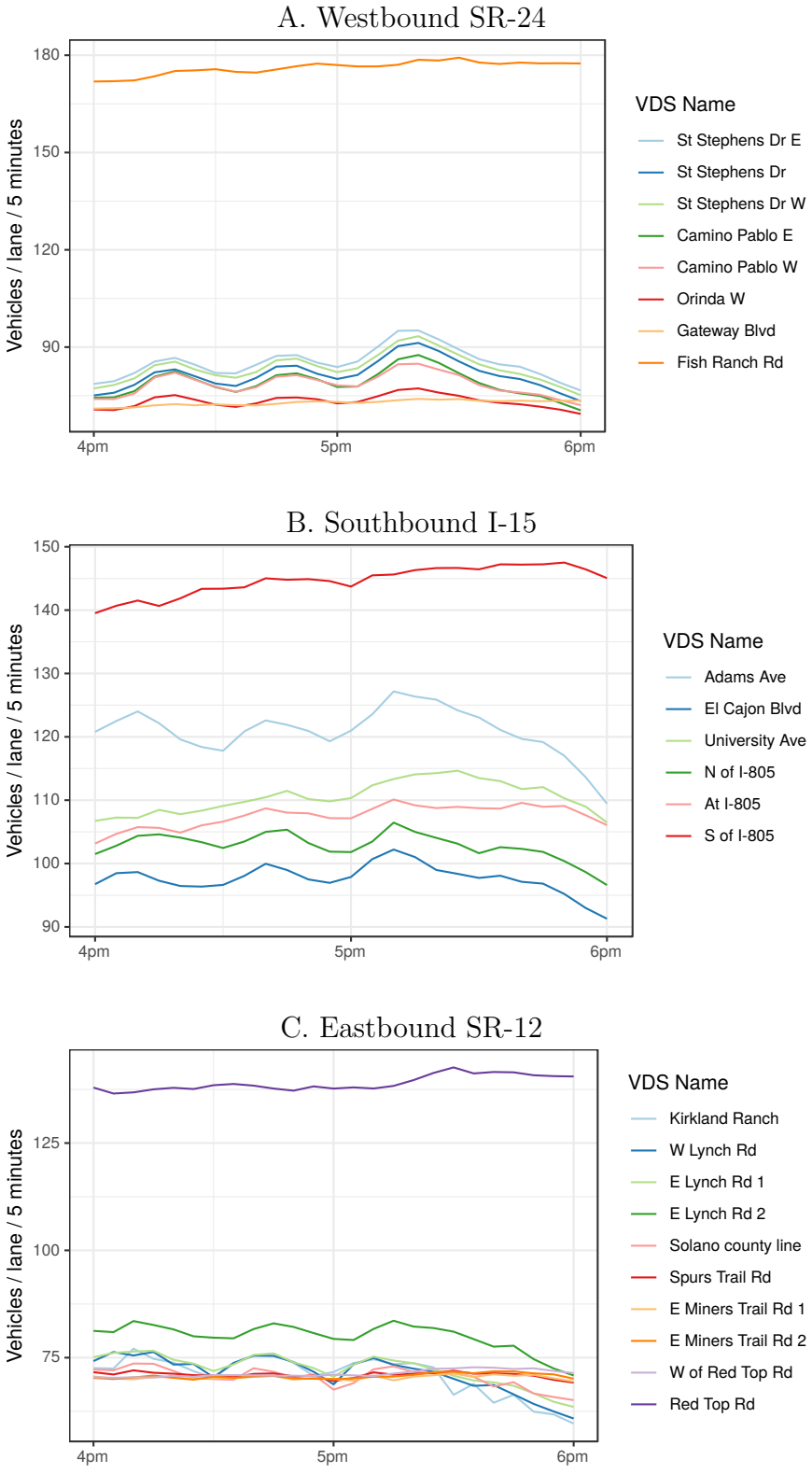


C. Eastbound SR-12



Notes: We exclude weekends and holidays. We define a queue as traffic moving under 30 miles-per-hour.

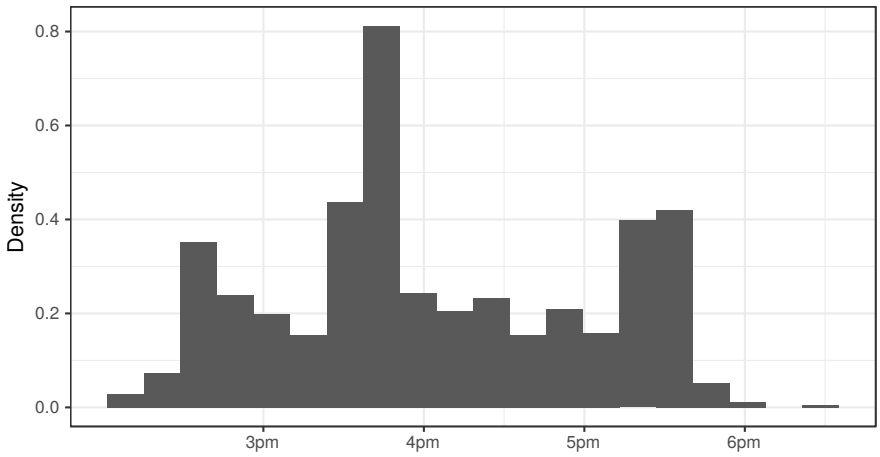
Figure A4: Mean Vehicle Flow on Weekdays When a Queue is Present



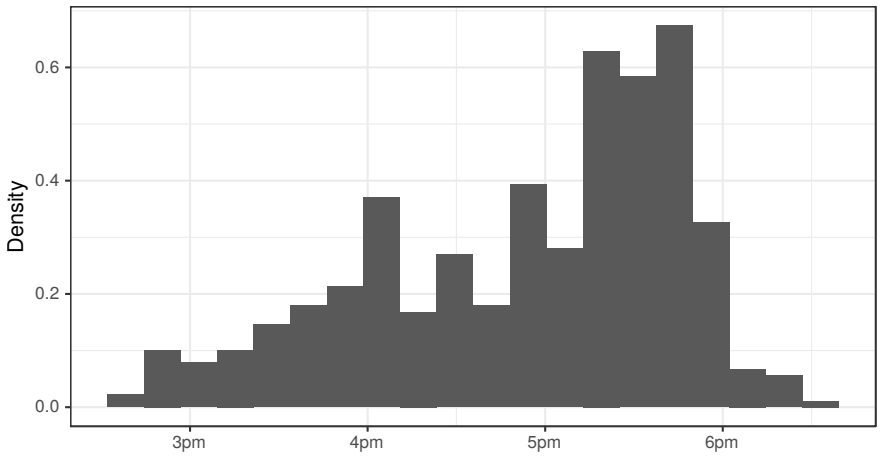
Notes: We exclude weekends and holidays. We define a queue as traffic moving under 30 miles-per-hour.

Figure A5: Time-of-Day that the Queue Begins Each Day, Histogram

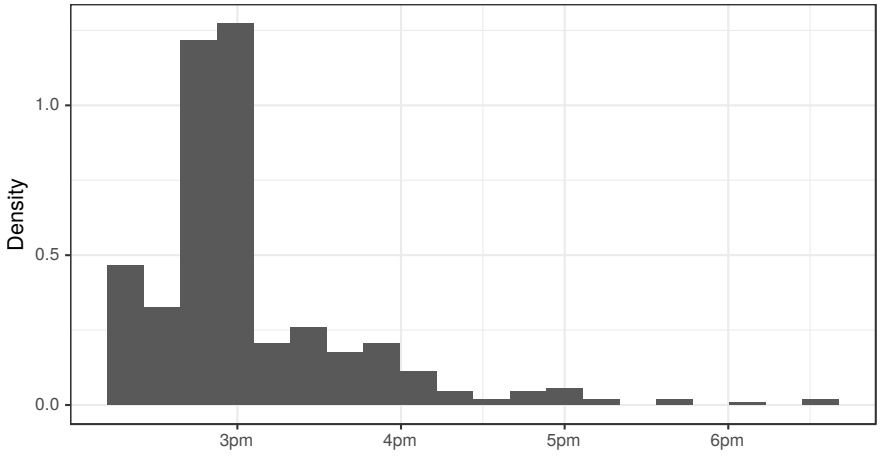
A. Westbound SR-24



B. Southbound I-15



C. Eastbound SR-12



Notes: For each day, we select the longest continuous period of time with a queue, and then we define the start of the queue as the beginning of that period.

Table A1: Summary Statistics, Complete Sample

SR-24	N	Mean	Std Dev	Min	Max
Year	287,272	2007.51	1.5	2005	2010
Queue present	287,272	0.18	0.4	0.00	1.00
Queue length (1000s of feet)	287,272	0.49	1.4	0.00	15.68
Capacity downstream of bottleneck (vehicles per 5 min per lane)	287,272	106.50	59.0	0.00	204.50
Speed upstream of bottleneck (mph)	287,106	53.29	20.4	3.00	80.50
Speed downstream of bottleneck (mph)	287,272	55.20	9.8	3.00	75.10

I-15	N	Mean	Std Dev	Min	Max
Year	276,636	2016.67	1.2	2015	2019
Queue present	276,636	0.05	0.2	0.00	1.00
Queue length (1000s of feet)	276,636	0.22	1.1	0.00	12.81
Capacity downstream of bottleneck (vehicles per 5 min per lane)	276,636	77.12	46.1	0.00	189.67
Speed upstream of bottleneck (mph)	276,611	64.01	11.2	3.00	76.00
Speed downstream of bottleneck (mph)	276,636	63.98	7.4	3.00	78.80

SR-12	N	Mean	Std Dev	Min	Max
Year	85,866	2018.19	0.4	2017	2019
Queue present	85,866	0.17	0.4	0.00	1.00
Queue length (1000s of feet)	85,866	0.99	2.6	0.00	19.55
Capacity downstream of bottleneck (vehicles per 5 min per lane)	85,866	69.67	51.7	0.00	210.00
Speed upstream of bottleneck (mph)	85,864	55.85	21.7	3.00	79.20
Speed downstream of bottleneck (mph)	85,866	54.86	10.7	3.00	75.00

Notes: The sample consists of all five-minute periods with non-missing data. For SR-24, data are from June 1, 2005 to June 30, 2010. For I-15, data are from January 1, 2015 to March 31, 2019. For SR-12, data are from January 1, 2017 to March 31, 2019. Weekends and holidays are excluded. We define a queue as present at a detector during a five-minute period if the average speed over that detector falls below 30 miles-per-hour. Speed upstream of bottleneck is measured at the first upstream detector (Gateway Boulevard, At I-805, and West of Red Top Road, respectively). Speed and capacity downstream of bottleneck are measured at the first downstream detector (Fish Ranch Road, South of I-805, and Red Top Road, respectively). Speeds and capacities are averaged across all lanes. Queue length is computed as the sum of non-missing segments between each detector and its downstream neighbor. For queue length only, we use PeMS-interpolated speeds if up to one lane is missing.

Table A2: Summary Statistics, 4–7pm, with Queue

SR-24	N	Mean	Std Dev	Min	Max
Year	26,159	2007.48	1.6	2005	2010
Queue length (1000s of feet)	26,159	3.20	2.5	0.00	15.68
Capacity downstream of bottleneck (vehicles per 5 min per lane)	26,159	176.20	12.2	0.00	204.50
Speed upstream of bottleneck (mph)	26,145	10.73	5.1	3.00	30.00
Speed downstream of bottleneck (mph)	26,159	41.71	4.2	5.85	54.00

I-15	N	Mean	Std Dev	Min	Max
Year	11,246	2016.71	1.3	2015	2019
Queue length (1000s of feet)	11,246	4.55	2.8	0.00	12.81
Capacity downstream of bottleneck (vehicles per 5 min per lane)	11,246	144.71	13.8	61.67	178.33
Speed upstream of bottleneck (mph)	11,245	17.80	4.7	3.60	30.00
Speed downstream of bottleneck (mph)	11,246	42.97	12.1	8.50	68.90

SR-12	N	Mean	Std Dev	Min	Max
Year	9,750	2018.19	0.4	2017	2019
Queue length (1000s of feet)	9,750	6.77	3.7	0.00	19.55
Capacity downstream of bottleneck (vehicles per 5 min per lane)	9,750	139.15	13.3	0.00	180.00
Speed upstream of bottleneck (mph)	9,748	9.16	2.0	3.20	29.90
Speed downstream of bottleneck (mph)	9,750	41.88	6.3	3.00	52.50

Notes: The sample includes all five-minute weekday periods between 4pm and 7pm when a queue is detected at the first upstream detector. For SR-24, data are from June 1, 2005 to June 30, 2010. For I-15, data are from January 1, 2015 to March 31, 2019. For SR-12, data are from January 1, 2017 to March 31, 2019. We define a queue as present at a detector during a five-minute period if the average speed over that detector falls below 30 miles-per-hour. Speed upstream of bottleneck is measured at the first upstream detector (Gateway Boulevard, At I-805, and West of Red Top Road, respectively). Speed and capacity downstream of bottleneck are measured at the first downstream detector (Fish Ranch Road, South of I-805, and Red Top Road, respectively). Speeds and capacities are averaged across all lanes. Queue length is computed as the sum of non-missing segments between each detector and its downstream neighbor. For queue length only, we use PeMS-interpolated speeds if up to one lane is missing.

Table A3: The Effect of Queue Length on Highway Capacity, OLS

	(1) Weekdays 4-7pm	(2) Weekdays 4:30-6:30pm	(3) Weekdays 4:30-6:30pm June-Aug
		A. Westbound SR-24	
Queue Length	-0.251	-0.250	0.529
	(0.153)	(0.145)	(0.139)
Total Observations	26,159	19,024	5,285
Number of Days	952	944	268
Dependent Variable Mean	176.2	177.1	177.8
Queue Length Mean	3.20	3.38	2.97
		B. Southbound I-15	
Queue Length	-0.431	-0.385	-0.081
	(0.115)	(0.110)	(0.160)
Total Observations	7,718	6,414	1,616
Number of Days	687	660	160
Dependent Variable Mean	149.2	149.5	153.6
Queue Length Mean	3.96	3.99	3.67
		C. Eastbound SR-12	
Queue Length	-0.105	-0.060	0.017
	(0.115)	(0.130)	(0.439)
Total Observations	9,750	6,803	1,335
Number of Days	297	295	58
Dependent Variable Mean	139.2	139.6	139.8
Queue Length Mean	6.77	7.27	6.60

Notes: This table reports estimates and standard errors from nine separate regressions, all estimated using ordinary least squares (OLS). The dependent variable in all regressions is highway capacity, measured in vehicles per five minutes per lane, downstream of the bottleneck. Queue length is measured in thousands of feet. The sample includes all five-minute periods with a queue. Standard errors are clustered by day.

Table A4: Alternative Specification with 25-Mile-Per-Hour Threshold, 2SLS

	(1) Weekdays 4-7pm	(2) Weekdays 4:30-6:30pm	(3) Weekdays 4:30-6:30pm June-Aug
		A. Westbound SR-24	
Queue Length	3.636	2.027	2.236
	(0.387)	(0.411)	(0.488)
Total Observations	25,075	18,340	5,035
Number of Days	934	924	261
Dependent Variable Mean	176.4	177.2	178.2
Queue Length Mean	3.18	3.35	2.95
K-P F-stat	78.1	68.7	24.5
		B. Southbound I-15	
Queue Length	3.356	4.137	0.822
	(1.018)	(1.117)	(0.754)
Total Observations	6,802	5,681	1,454
Number of Days	653	625	152
Dependent Variable Mean	149.1	149.4	153.6
Queue Length Mean	3.37	3.39	3.21
K-P F-stat	18.5	12.0	4.41
		C. Eastbound SR-12	
Queue Length	0.073	-0.179	-0.695
	(0.195)	(0.292)	(0.628)
Total Observations	9,700	6,783	1,330
Number of Days	297	295	58
Dependent Variable Mean	139.2	139.6	139.8
Queue Length Mean	6.66	7.14	6.51
K-P F-stat	114.1	71.1	31.1

Notes: This table reports results from an alternative specification in which everything is identical to the specification used for our baseline results in Table 1, except we assume that a queue is present whenever traffic is moving at under 25 miles-per-hour (rather than 30 miles-per-hour).

Table A5: Alternative Specification with 35-Mile-Per-Hour Threshold, 2SLS

	(1) Weekdays 4-7pm	(2) Weekdays 4:30-6:30pm	(3) Weekdays 4:30-6:30pm June-Aug
		A. Westbound SR-24	
Queue Length	3.492	1.921	2.501
	(0.341)	(0.384)	(0.499)
Total Observations	27,054	19,592	5,455
Number of Days	973	965	276
Dependent Variable Mean	176.1	176.9	177.7
Queue Length Mean	3.23	3.42	3.00
K-P F-stat	91.4	78.2	27.5
		B. Southbound I-15	
Queue Length	1.831	2.184	-0.142
	(0.465)	(0.490)	(0.395)
Total Observations	8,379	6,932	1,729
Number of Days	717	690	164
Dependent Variable Mean	149.3	149.6	153.6
Queue Length Mean	4.59	4.63	4.30
K-P F-stat	35.4	25.8	10.1
		C. Eastbound SR-12	
Queue Length	0.112	-0.149	-0.758
	(0.185)	(0.265)	(0.601)
Total Observations	9,796	6,826	1,341
Number of Days	297	295	58
Dependent Variable Mean	139.1	139.6	139.8
Queue Length Mean	6.90	7.40	6.72
K-P F-stat	119.7	75.9	27.6

Notes: This table reports results from an alternative specification in which everything is identical to the specification used for our baseline results in Table 1, except we assume that a queue is present whenever traffic is moving at under 35 miles-per-hour (rather than 30 miles-per-hour).

Table A6: Alternative Event Study Analyses, Median Regressions

	(1) Ten Minute Window	(2) Twenty Minute Window	(3) Thirty Minute Window
		A. Westbound SR-24	
	1.500	1.750	2.000
	(0.333)	(0.287)	(0.243)
In-window median	171	171.5	171.5
Number of days	706	706	706
		B. Southbound I-15	
	1.333	1.500	-0.333
	(0.545)	(0.475)	(0.408)
In-window median	146	146.3	147.3
Number of days	694	694	694
		C. Eastbound SR-12	
	-2.000	-2.500	-2.667
	(1.233)	(0.838)	(0.704)
In-window median	136	136	136
Number of days	247	247	247

Notes: This table reports results from alternative event study analyses in which everything is identical to the specification used in Table 2, except we use median regressions.