Journal of Public Economics 187 (2020) 104197

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

An empirical test of hypercongestion in highway bottlenecks☆

Michael L. Anderson, Lucas W. Davis *

University of California, Berkeley, United States of America

ARTICLE INFO

Article history:
Received 16 October 2018
Received in revised form 13 April 2020
Accepted 19 April 2020
Available online 28 May 2020

Keywords:
Hypercongestion
Traffic congestion
Capacity drop
Speed
Traffic flows

JEL:
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 in the context of highway bottlenecks. We propose an empirical test using an event study design to measure changes in highway capacity at the onset of queue formation. We apply this test to three highway bottlenecks in California for which detailed data on traffic flows and vehicles speeds are available. We find no evidence of a reduction in highway capacity at any of the three sites during periods of high demand. Across sites and specifications we have sufficient statistical power to rule out even small reductions in highway capacity. This lack of evidence of hypercongestion stands in sharp contrast to most previous studies and informs core models in urban and transportation economics.

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.

Our paper revisits this idea in the context of highway bottlenecks. We propose an empirical test using an event study design to measure changes in highway capacity at the onset of queue formation. Event study designs have become ubiquitous in empirical microeconomics and finance (see, e.g. Duggan et al., 2016; Dobkin et al., 2018; Freyaldenhoven et al., 2019), but they are novel in our context. As we discuss later, event study designs have several advantages relative to the empirical approaches used in the existing literature.

We apply our empirical test to three highway bottlenecks in California. For each study site, we observe highly-detailed data on traffic flows and vehicle speeds at several locations before and after the bottleneck. Although the three sites have different features, all have bottlenecks that generate long queues during weekday afternoons.

We find no evidence of hypercongestion at any of the three sites. Vehicle speeds decrease sharply from above 50 miles-per-hour to below 20 miles-per-hour at the onset of queueing. However, we find that the rate at which vehicles flow through the bottleneck, measured in vehicles per five minutes, is essentially constant throughout the period of queue formation. Results are similar across all three sites and a range of alternative specifications, with no discernible reduction in highway capacity during periods of high demand.

This lack of evidence of hypercongestion stands in sharp contrast to most previous studies. 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) all find evidence of "capacity drop", "flow breakdown", or the "two capacity phenomenon" at bottlenecks, referring to a drop in roadway capacity upon queue formation.

☆ 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.

* Corresponding author.
E-mail addresses: mlanderson@berkeley.edu (M.L. Anderson), lwdavis@berkeley.edu (L.W. Davis).

https://doi.org/10.1016/j.jpubeco.2020.104197
0047-2727/

The absence of evidence of hypercongestion is not due to a lack of statistical precision. Whereas many previous studies use data only from a single day or small handful of days, our event study approach aggregates information from 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.

Our findings directly inform several core models and concepts in urban and transportation economics. The bottleneck model is a canonical model in the field (Vickrey, 1969; Small, 1982; Arnott et al., 1990, 1993, 1994). In this model drivers face a tradeoff between time delays and schedule inflexibility and optimize their departure times accordingly. The question of whether bottlenecks have variable or constant capacity has important welfare and policy implications, but it is fundamentally an empirical question, which our study answers.

We also discuss the policy implications of our results. 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. Arnott et al., 1993). Thus whether or not hypercongestion exists likely has minimal impact on the how taxes are set in the optimal Pigouvian solution. Without hypercongestion the welfare gains from optimal congestion pricing are smaller, however, as total social costs are lower in the unregulated equilibrium.

Our paper is germane to a growing empirical literature on the formation of traffic congestion. For example, 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. (forthcoming) uses variation from driving restrictions to estimate the marginal external cost of traffic congestion in Beijing.[1] Russo et al. (2019) uses public transportation strikes as an instrument for traffic density in estimating the marginal external cost of traffic congestion in Rome.[2] Akbar and Duranton (2017) uses travel surveys and other data from Bogotá, Colombia to estimate the deadweight loss of traffic congestion.[3]

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 in the United States (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. Background

2.1. Conventional wisdom regarding hypercongestion

It is clear that traffic congestion reduces speed. But there is a widespread view among transportation engineers and economists that as demand for travel goes up, this decreases not only speed but also the *capacity* of the road system. There are two primary forms of hypercongestion, both involving bottlenecks.

In the first form of hypercongestion, there is a "spillover" from one bottleneck to other routes. This occurs when the queue behind a bottleneck grows so long that it blocks or impedes some other route. These "queue spillovers" or "triggerneck" situations are particularly prevalent in dense urban networks, with gridlock as an extreme example, but they can also occur on highways, for example when a queue on a highway backs up far enough to block upstream exits. Several economic analyses have examined hypercongestion in such contexts, often with an emphasis on dense urban networks (see, e.g. Small and Chu, 2003; Arnott, 2013; Fosgerau and Small, 2013; Small, 2015). Arnott (2013), for example, proposes a "bathtub" model of hypercongestion for downtown areas in which capacity decreases at high levels of traffic density. As we previously noted, our analysis and results do not speak to this type of hypercongestion.

In the second form of hypercongestion, the capacity of the road system decreases at the onset of queue formation. Unlike the first form of hypercongestion, this second form of hypercongestion does not require there to be multiple bottlenecks, nor for there to be "spillovers" of any kind across routes. Instead, the idea is that bottlenecks intrinsically have two different capacity levels, one when there is no queue, and then another, lower capacity level, after a queue has formed. Accordingly, the literature has sometimes referred to this form of hypercongestion as the "two capacity phenomenon" or "capacity drop". This decrease in capacity is in addition to the standard externality caused by the lengthening of the queue, and the literature has generally been clear that this capacity drop refers to a change in traffic flows, measured in vehicles per unit of time, crossing through the bottleneck.

This capacity drop feature of bottlenecks is viewed as firmly established in the transportation engineering literature. For example, the first sentence of Jin et al. (2015) reads, "Since the 1990s, the so-called two-capacity or capacity-drop phenomenon of active bottlenecks, in which 'maximum flow rates decrease when queues form', has been observed and verified at many bottleneck locations." Yuan et al. (2015) explains "Traffic jams reduce the capacity of the road. This phenomenon is called the capacity drop. Because of capacity drop, traffic delays increase once congestion sets in." Leclercq et al. (2016) writes, "Effective capacity is referred [to] in some papers as the queue discharge rate. Experimental findings show that capacity drops are often observed at merges even if downstream traffic conditions are in free-flow. The magnitude of the capacity drops is mentioned to be between 10% and 30% of the maximal observed flow." And from Lamotte et al. (2017), "Indeed, most real-world bottlenecks have reduced passing rates [i.e. capacity] for highly congested conditions. This phenomenon is known in transportation economics as hypercongestion."[4]

A growing economics literature explores the policy and welfare implications of the capacity drop phenomenon. Most recently, it appears in

[1] Beijing's driving restrictions are based on the last digit of the license plate and only 2% of vehicles have a license plate ending in "4". Yang et al. (forthcoming) use this as an instrument for traffic flows, finding that the marginal external cost of traffic congestion is $0.30 per vehicle-kilometer. In their empirical analysis they focus on ordinary congestion, but highlight hypercongestion as a key priority for future research.

[2] Public transportation strikes are common in Rome, and Russo et al. (2019) use strikes as well as hour-of-week fixed effects to instrument for traffic density. They estimate that the marginal external cost of road congestion is $0.22 per vehicle-kilometer, with about one-fourth of these costs borne by bus travelers.

[3] 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.

[4] 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 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".

a pair of innovative papers by economist Jonathan Hall. These papers apply hypercongestion to a bottleneck model and show that highway pricing can generate a Pareto improvement when agents are heterogeneous, even before redistributing toll revenues (Hall, 2018, forthcoming). Motivated by both forms of hypercongestion (queue spillovers and capacity drop), these papers use a model in which highway capacity drops by 10% or more once a queue develops. For example, Hall (forthcoming), Table 5, reports welfare effects of congestion pricing for capacity drops of 10%, 17.5%, and 25%. Central to these analyses is the use of "Lexus Lanes", i.e. a subset of lanes that are tolled while others remain free. Pricing these lanes can increase capacity by eliminating queues, and the remaining free lanes provide an option to inflexible, lower-income drivers.

Most papers attribute the capacity drop to lane-changing behavior. Before a queue forms, motorists at a merge are better able to fill in gaps between vehicles and use all available highway capacity. However, once a queue forms, vehicles must slow down considerably or even come to a complete stop before merging. When a motorist merges in after a previous vehicle, they often leave a gap between vehicles. If they are not able to accelerate quickly to fill the gap, this space ends up being lost capacity. In addition, when there is a queue motorists often perform what transportation engineers refer to as a "destructive lane change", which means they force their way into the other lane while moving slowly, often leaving a gap in front of them.[5]

2.2. Empirical studies of capacity drop

Table 1 summarizes the existing empirical literature on capacity drop at highway bottlenecks. All 14 studies that we reviewed find evidence of a capacity drop. Estimates range as high as 16.3%, and the median capacity drop is about 10%. Hall (2018) performs a similar review of this literature, reporting that 16 out of 17 papers find evidence of a capacity drop, with estimates ranging as high as 25%, and a median capacity drop of 10%.[6]

This literature has been widely read and is influential. For example, the papers in Table 1 have been cited over 2700 times, collectively, according to Google Scholar. In this section we describe several of the studies in more detail. The 14 studies use a variety of different study sites, data sources, and empirical approaches. We explain why this setting is particularly challenging for making causal statements, and we point to several recurring identification concerns which motivate our empirical analyses.

One of the first and most influential studies is Banks (1990). Using an approach that is typical in the broader literature, Banks (1990) plots nine days of data on traffic flows on I-8 in San Diego. It then uses "visual inspection" of detector data and videotapes to mark the moment of queue formation, based upon a heuristic combination of speeds, vehicle spacing, and lane use. It finds an average decrease in flows of 2.8% at the onset of queue formation. Banks (1990) describes this as the "two capacity phenomenon", evoking the idea that highways have one capacity when there is no queue, and then another, lower capacity, after a queue has formed.

Another influential study in this literature is Persaud et al. (1998). This paper measures capacity drops at multiple sites in Toronto, finding capacity drops ranging from 10.6% to 15.3%. Like Banks (1990), this paper uses visual inspection of speeds and flows to identify the exact moment of queue formation. It also uses visual inspection to determine the exact time period over which the flow average is calculated, with a view toward selecting a pre-queue period with an unusually high flow level.[7]

A potential concern with these analyses is selection bias on the part of the researcher. Traffic flows vary widely from minute to minute. For example, some vehicles are driven faster than others. Consequently, an approach based on visual inspection of flows risks attributing to capacity drop what may actually be high-frequency variability in flows. Said differently, when presented with a noisy time series on traffic flows it is relatively easy for a researcher to find moments in which flows decrease suddenly, but this is not the same as identifying the causal impact of queueing. Neither Banks (1990) or Persaud et al. (1998) have a direct measure of queueing, so they approach causality from the other direction, looking for a moment in time when traffic flows decrease, and then inferring that a queue formed in that moment.

Selection bias can occur in subtle ways. For example, Zhang and Levinson (2004) reports capacity drops ranging from 2% to 11% based on data from multiple bottlenecks in the Twin Cities area in Minnesota. They use density thresholds to determine whether locations are congested or uncongested. However, they then use visual inspection to determine which periods to include when calculating the pre-queue average flow. Like Persaud et al. (1998), they explicitly select pre-queue periods with unusually high flow levels.[8] Again, the concern with selecting a pre-queue period with unusually high flow levels is that it may introduce selection bias; average flows will tend to decrease due to mean reversion following an interval conditioned on having abnormally high flows.

Later studies that emphasize cumulative vehicle counts (Bertini and Leal, 2005; Cassidy and Rudjanakanoknad, 2005) are subject to similar concerns about selection bias. By plotting cumulative vehicle counts

Table 1
Previous studies of capacity drop at highway bottlenecks.

Paper	Capacity Drop (%)	Location
Banks (1990)	2.8	I-8, San Diego
Hall and Agyemang-Duah (1991)	5.8	Queen Elizabeth Way, Toronto
Banks (1991)	−1.2 to 3.2	Multiple Sites, San Diego
Persaud et al. (1998)	10.6 to 15.3	Multiple Sites, Toronto
Cassidy and Bertini (1999)	7.4 to 8.7	Multiple Sites, Toronto
Bertini and Malik (2004)	4.0	US-169, Minneapolis
Zhang and Levinson (2004)	2.0 to 11.0	Multiple Sites, Twin Cities, MN
Bertini and Leal (2005)	9.7	M4, London
	12.0	I-494, Minneapolis
Cassidy and Rudjanakanoknad (2005)	11.7	I-805, San Diego
Chung et al. (2007)	12.3	I-805, San Diego
	6.2	SR-24, San Francisco Bay Area
	5.8	Gardiner Expressway, Toronto
Guan et al. (2009)	15.0	Fourth Ring Road, Beijing
Oh and Yeo (2012)	8.9 to 16.3	Multiple Sites, California
Srivastava and Geroliminis (2013)	15.0	US-169, Minneapolis
Jin et al. (2015)	10.5	I-405, Irvine, CA

Notes: Similar tables appear in Oh and Yeo (2012) and Hall (2018).

[5] Hall (2018) explains that once a queue forms, vehicles "need to change lanes" and that "when traffic is heavy, doing so is difficult; there will typically be a vehicle that comes to a stop before merging and, rather than waiting for a gap, will force its way over." Similarly, Srivastava and Geroliminis (2013) attributes the capacity drop to, "lane changing maneuvers, vehicles entering a merge at slow speeds, and heterogeneous lane behavior". Leclercq et al. (2016) explains, "The main physical explanations for such a phenomenon are lower speeds for merging vehicles combined with bounded acceleration, and the impacts of driver behaviors. In a nutshell, slower vehicles create voids in front of them that locally reduce the available capacity and lead to temporal flow restrictions."

[6] The one paper reviewed by Hall (2018) that does not find evidence of capacity drop is Hurdle and Datta (1983), a somewhat older paper that is not focused explicitly on capacity drop but that includes figures describing traffic flows before and after a queue forms on three mornings in May 1977 at a highway bottleneck near Toronto, Canada.

[7] Specifically, Persaud et al. (1998) explains that the beginning of the pre-queue period, T_d, was selected explicitly so that the pre-queue flow average would be systematically higher than the post-queue flow average (Q_d). From p. 65, "Once again, visual inspection was employed. T_d was taken as the time at which Q_d was continually exceeded in the pre-queue period."

[8] From p. 126, "Therefore, τ_s [the beginning of the pre-queue period] is determined by the interval in which the flow at a freeway section exceeds its long-run queue discharge flow."

from multiple detectors it becomes possible to see queues emerge, visible as a reduction in flow at further downstream detectors relative to upstream detectors. While initially this might appear to mitigate the problem of selection bias, it actually suffers from identical concerns. In particular, it continues to be difficult to separate capacity drop from the usual minute-to-minute variability in traffic flows.[9] In both cases a researcher uses visual inspection based in part on the dependent variable to infer when a queue forms.

Selection bias can occur even when researchers use alternatives to visual inspection. For example, Oh and Yeo (2012) critiques previous studies on the basis that the "visual inspection" approach is "arbitrary", but then proceeds to measure pre-queue flow using the "maximum 5-minute flow before bottleneck activation was observed." (p.115) Even without visual inspection, this approach can still introduce selection bias because average flows will tend to decline following a period selected to have unusually high flows.[10] As with the other studies, the fundamental challenge is that traffic flows are highly variable, so any *ex post* selection based in part or whole on this variable can lead a researcher to mechanically find evidence of capacity drop due to mean reversion.

3. Our empirical test

In this section we describe our empirical test of whether highway capacity decreases when a queue forms. Our test takes the form of a standard event study regression.

The test is designed to be applied in highway settings with a single bottleneck — locations where some physical feature of the highway serves to restrict 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, queue spillovers, and alternative routing opportunities.

The event of interest in our context is the moment in time that the queue forms. How we define and measure queue formation is critical for our analysis, but we defer that discussion until later (Section 4.5), after introducing the study sites.

The event study regression allows us to assess whether there is a change in highway capacity at the onset of a queue. In particular, we estimate regressions of the form:

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

The dependent variable in these regressions is traffic flow in 5-min 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 min) before the queue forms, $\tau = 16$ for 16 periods (i.e. 80 min) 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 (or, in the case of median regressions, least absolute deviations), it is equivalent to taking conditional averages in event time. Some event studies drop the indicator for $\tau = -1$ to avoid perfect collinearity, but we instead suppress the regression intercept. This choice does not affect inference, but it enables us to easily generate figures mapping out traffic flows or traffic speeds in event time. We cluster our standard errors by date, allowing for arbitrary serial correlation in the dependent variable within a day.[11]

The event study analysis focuses on the transition between no queue and queue. We do not restrict the sample to include only observations in which there is a queue, as that would omit observations before $\tau = 0$. Nevertheless, in our empirical applications we tend not to see large increases in flow leading up to queue formation, suggesting that flow is near capacity for an extended period of time prior to queue formation, and we refer to the dependent variable in these regressions as capacity, rather than flow.

Before introducing our study sites, we highlight three advantages of the event study approach relative to the empirical approaches used in the existing capacity drop literature (Section 2.2).

First, the event study provides a natural approach for aggregating information from multiple days. In contrast, many previous studies examine data one day at a time and must contend with minute-to-minute variability in traffic flows. Aggregating across hundreds of days reduces the influence of minute-to-minute fluctuations, reducing the risk of spurious findings and increasing statistical precision.

Second, the event study approach forces us to adopt an objective, standardized rule for identifying the moment of queue formation. Whereas previous studies use visual inspection or other ad hoc approaches, we can estimate and demonstrate a "first-stage" relationship directly, and identify queue formation based on measuring traffic speeds — and not flows — thereby mitigating concerns about selection bias.

Third, the event study approach lends itself well to statistical inference. In the capacity drop literature, few studies report standard errors, and we were not able to find a single study that reports standard errors that account for serial correlation. In contrast, it is straightforward with the event study regression in Eq. (1) to construct confidence intervals and perform formal statistical tests that account for potential dependence in the errors.

Despite its strengths, our event study approach also has limitations. In particular, while we focus our analysis on times of day when a queue typically forms due to high demand, we cannot rule out the possibility that some queues may form in response to roadway incidents that restrict capacity. Our event study analysis thus could have some bias toward finding capacity drops — reverse causality might result in a roadway capacity drop generating a queue, rather than vice versa. We therefore view our estimates as upper bounds on the magnitude of capacity drop at our study sites.

4. Empirical application

We apply our empirical test using data from three study sites. All three sites are in California, allowing us to use high-quality, comparable data from a single source, the California 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.[12] In this section we describe the study sites

[9] For example, Bertini and Leal (2005) use data from a single day of traffic on the M4 in London, and a single day of traffic on the I-494 in Minneapolis. Plotting cumulative vehicle counts for consecutive traffic detectors, they use visual inspection to determine the moment of queue formation, and then visual inspection to determine the exact time periods to use for calculating pre- and post-queue flow averages (the slope in cumulative vehicle counts). On the M4, for example, they mark the queue's start at 6:45 a.m., noting, "Excess vehicle accumulations occurred between [upstream] Detectors 6 and 7 subsequent to flow reductions observed at [downstream] Detectors 7 and 8 around 6:44 and 6:45 a.m., respectively." (p. 399) Since the dependent variable is discharges from downstream detectors (7 and 8), it is unsurprising that they find evidence of a capacity drop.

[10] In a related example, Hall and Agyemang-Duah (1991) uses statistical significance in flow differences as a factor in deciding when capacity drop has occurred. Although this rule may be less arbitrary, the approach still introduces selection bias because it leads the researcher to focus on a non-random subset of periods in which large decreases occurred.

[11] Serial correlation *across* days is not a concern for standard errors because the independent variables, by construction, are perfectly balanced (i.e. uncorrelated) across days.

[12] 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/.

(Section 4.1) and present descriptive statistics on traffic flows (Section 4.2) and vehicle speeds (Section 4.3) before turning to measuring the onset of the queue (Section 4.5).

4.1. Site selection

We selected 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. We dropped several promising sites because loop detectors were not available. We have not performed a comprehensive survey of all potential sites in California, but with over 380,000 total lane-miles of highway in the state and nearly 40,000 installed loop detectors, there are almost certainly other study sites in California that would satisfy the criteria of having a clearly-identified single bottleneck and good data coverage.

4.1.1. Site 1

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. Traffic delays are common at this location; indeed, transportation engineers have repeatedly studied this exact site (Chin and May, 1991; Chung and Cassidy, 2002; 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.[13]

Fig. 1 depicts the study site. Approximately 3000 ft 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.[14] For westbound travelers there is no reasonable alternative to traversing the tunnel.[15]

4.1.2. Site 2

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 Fig. 1 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 Site 2 with results from the other two sites where there is less scope for substitution to alternative routes.[16]

4.1.3. Site 3

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 Fig. 1 illustrates, at this location SR-12 merges from two lanes down to one lane.[17] 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.

In summary, all three sites contain specific locations where the number of lanes decreases sharply. An alternative bottleneck type would have been one in which a highway entry ramp from a surface street or a highway junction merges into the highway lanes. Highway entry ramps are a common form of bottleneck, but also tend to result in less predictable queueing behavior than the locations we consider. As we show below, at our three sites queues form predictably almost every weekday afternoon, and once formed, tend to last for an hour or more. These features make our sites particularly amenable for empirical analysis. Nevertheless, it would be interesting in future work to apply our event study design to highway entry ramps.

4.2. Traffic flows

Fig. 2 plots average traffic 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 Site 1, but as noted earlier we focus on afternoons 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 average vehicle flows per lane that are approximately twice as high at

[13] 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.

[14] The first upstream detector is approximately 1000 ft from the bottleneck. This spacing introduces some delay between the formation of the queue and its detection. Detectors at other sites — in particular at Site 2 — are located closer to their respective bottlenecks. Reassuringly, the estimates from our event study analysis are similar across all three sites, suggesting that our results are not driven by the particular spacing of the detectors at any one particular site.

[15] For visual clarity the figure does not include exits and entrances. One of the 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 9000 ft (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.

[16] One advantage of the Site 2 site is that the first upstream detector, at I-805, is located only 300 ft from the bottleneck. This proximity means that any queue is detected almost immediately, since even emergency braking from freeway speeds requires up to 200 ft to stop.

[17] The first upstream detector, W of Red Top Rd, is located approximately 700 ft from the bottleneck. This spacing is closer than on Site 1 but further than on Site 2.

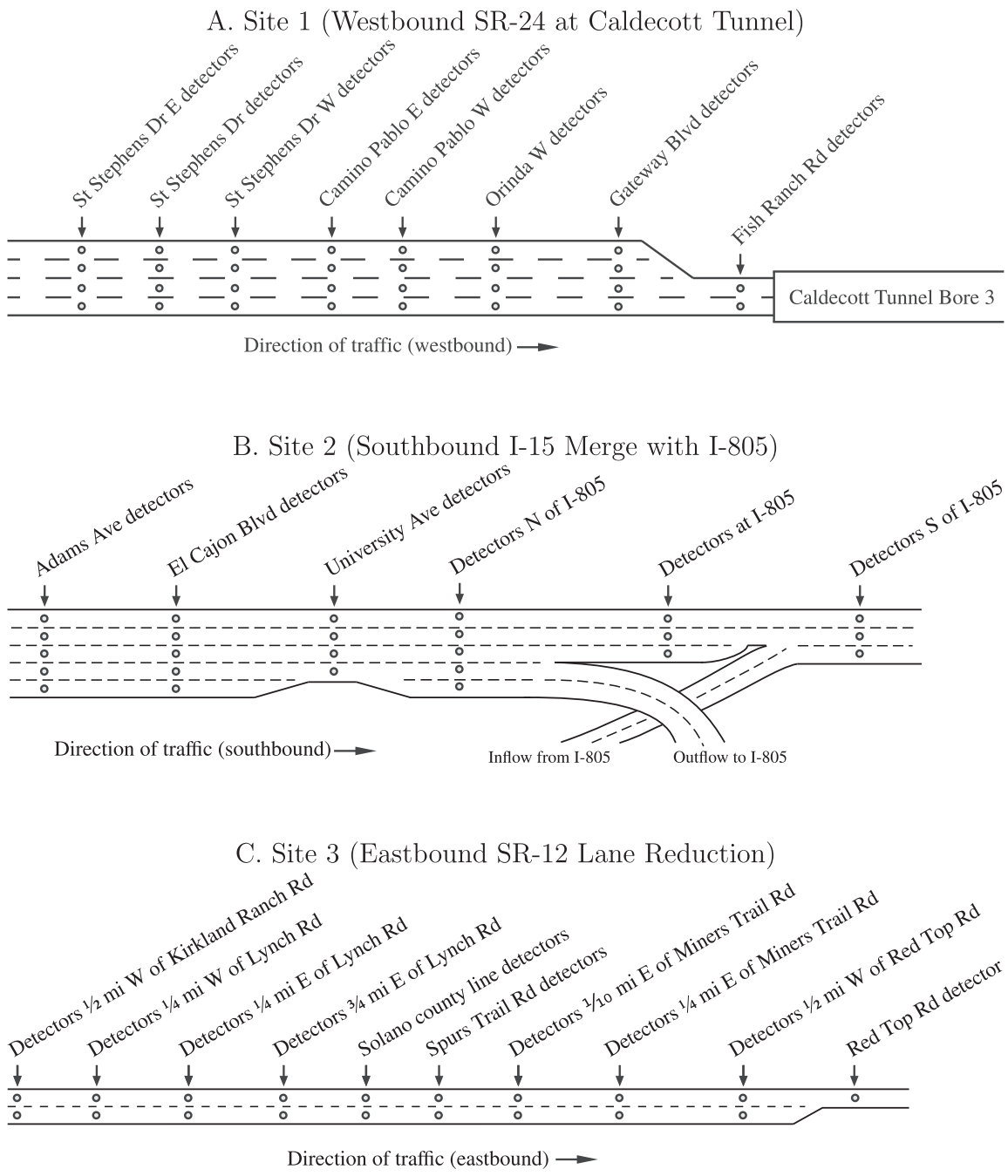

Fig. 1. Study sites. Notes: Figures approximately to scale.

the downstream location (Fish Ranch Rd) as compared to upstream locations.

At Sites 2 and 3, total vehicle traffic peaks in the afternoon. As with Site 1, 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 Site 3, 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.

4.3. Vehicle speeds

Fig. 3 plots average 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, at Site 1 the detector immediately upstream of the bottleneck (Gateway Blvd) exhibits average speeds below 40 miles-per-hour between about 3 pm and 6 pm. At Site 2 all six detectors experience large decreases in speed during afternoon hours. Finally, Site 3 has the most severe afternoon decreases in speed, with several upstream detectors exhibiting average speeds below 30, or even below 20, miles-per-hour.

Speeds tend to decrease much less at downstream detectors. At Site 1, for example, average speeds immediately upstream (Gateway) and downstream (Fish Ranch) track each other closely throughout most of the day. Between 3 pm and 6 pm, however, there is a significant divergence; upstream speeds slow to below 20 miles-per-hour, while

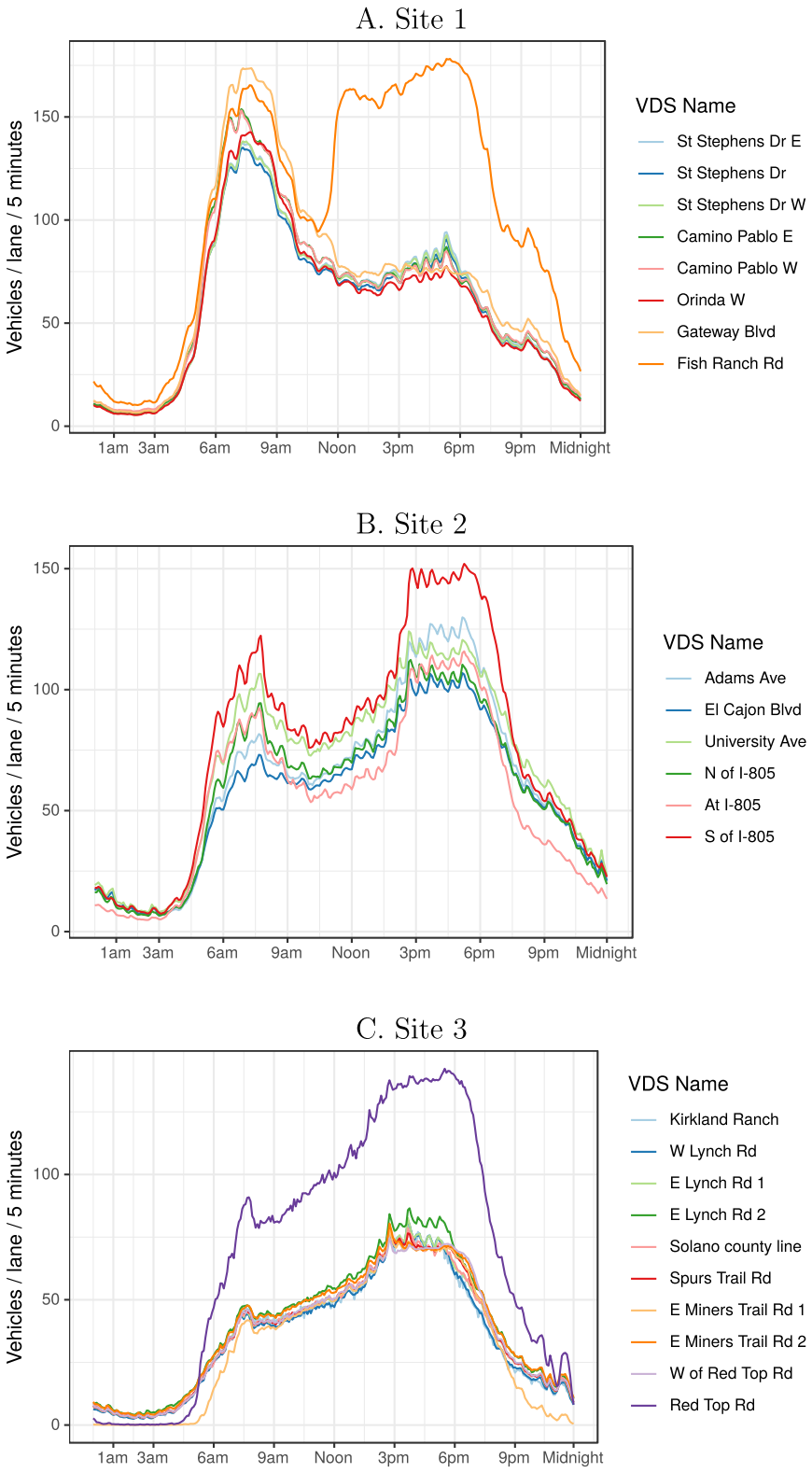

Fig. 2. Average traffic flows. Notes: We exclude weekends and holidays. Vehicle detector stations (VDS) are ordered in the direction of traffic, and for each site only the last station in the list is located downstream of the bottleneck.

downstream speeds remain above 40 miles-per-hour. Similarly, at Site 3, 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.

4.4. Conventional measures of capacity drop

Before defining the queue onset and estimating our main results, we reproduce the generic capacity-drop result from the existing literature.

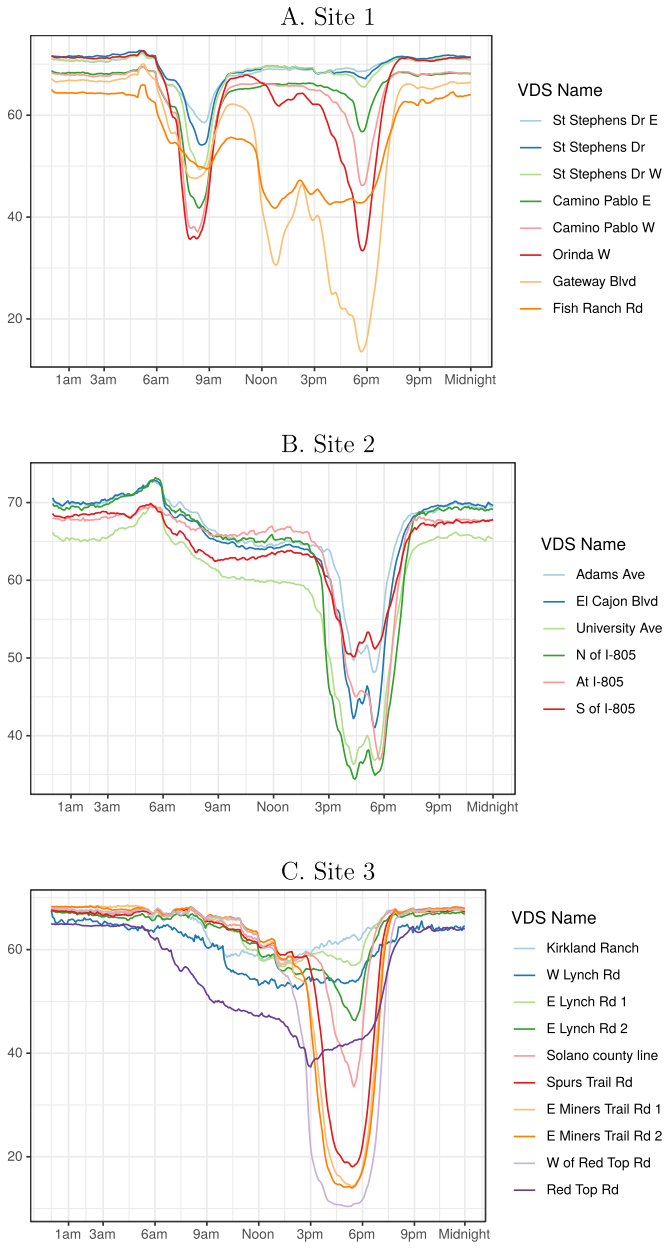

Fig. 3. Average vehicle speeds (in miles-per-hour). Notes: We exclude weekends and holidays.

To estimate the "conventional" capacity-drop model we compare observed flows for 10 min prior to queue formation with observed flows for 20 min following queue formation. For this exercise we follow Zhang and Levinson (2004), coding a queue as forming if average occupancy (the fraction of time that a detector has a vehicle above it) exceeds 25%.[18] In addition, we condition the sample to only contain queues for which the maximum observed flow rate in the 10-min period prior to queue formation exceeds the long-run (60-min) flow rate from the bottleneck by approximately 5%. This conditioning is similar to the sample-selection criteria used in several previous capacity-drop studies, including Persaud et al. (1998), Zhang and Levinson (2004), and Oh and Yeo (2012).

Appendix Table A1 reports results from this exercise. For each site, we choose a random sample of 50 days that meet the criteria described above, for 150 site-days in total. Columns (1) and (2) report the average traffic flows 10 min prior to queue formation and 20 min after queue formation, respectively.[19] Column (3) reports the change in traffic flows after queue formation, i.e. the difference between Columns (1) and (2). In all columns we normalize the measure to represent vehicle flows (or change in vehicle flows) per lane per five minutes, irrespective of site geometry or measurement window length. Negative changes indicate a decrease in capacity.

The vast majority of days at all sites reveal negative changes, implying capacity drops. The estimated magnitudes of the mean capacity drops are 5.3%, 6.8%, and 5.0% at Sites 1, 2, and 3 respectively. These values fall within the range reported in existing studies (see Table 1) and are similar in magnitude to the average capacity drop (5%) found in Zhang and Levinson (2004).

In summation, applying conventional capacity-drop models to our data reveals evidence of capacity drop at all three sites. Nevertheless, the sample-selection criteria used in many conventional models is, we believe, sensitive to mean reversion. The results in Appendix Table A1 suggest that, if our models generate different conclusions than conventional models, this divergence is due to differences in our approach for defining and measuring queues rather than differences in data types or bottleneck study sites.

4.5. Measuring the onset of the queue

We now turn to focus explicitly on the formation of the queue. Aggregate patterns of traffic flows and vehicle speeds imply that there is significant queuing of vehicles during afternoon hours at all three sites. With a mild assumption we can use our data to measure the presence of vehicle queues more directly. As a baseline, we assume that a queue is present whenever traffic is moving at under 30 miles-per-hour at the upstream detector closest to the bottleneck. This threshold is arbitrary, but we show that our results are robust to alternative definitions. This assumption provides an objective, standardized rule for determining whether a queue is present, and with this rule we determine the time each day when the queue initially forms.

Fig. 4 plots for each site the percentage of hours with a queue present, using our 30 miles-per-hour preferred threshold. During morning hours, there are almost never queues at any of the three sites. Then, during afternoon hours, queues become much more common. The exact pattern varies across sites, but by 6 pm there are queues during almost 100% of weekdays at Sites 1 and 3, and during about 50% of weekdays at Site 2. Queues then dissipate at all three sites between 7 pm and 8 pm, with almost no queueing after 9 pm at any site.

Fig. 5 presents for each site a histogram of the time-of-day at which the queue begins each day. For each day, we selected the longest continuous period of time with a queue, and we defined the start of the queue as the beginning of that period. Queues at all three sites tend overwhelmingly to begin between 2:30 pm and 6 pm. There is variation across sites and days. At Sites 1 and 2, the queue sometimes starts before 3 pm, but on many days does not start until after 5 pm. For Site 3 there is less variation, with the queue frequently starting between 2:30 pm and 3:00 pm.

We conduct subsequent analyses in "event time", or time in minutes relative to the onset of the queue. We normalize event time so that the longest-duration afternoon queue begins at time zero on each day. 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 afternoon peak hours

[18] We choose Zhang and Levinson (2004) as a template for several reasons. First, their loop-detector data are similar in nature to ours, and they cover a wide variety of sites. Second, their overall approach is broadly representative of strategies that a number of capacity drop studies have implemented. Third, their exact methodology is less ad hoc than some of the other studies, and relies less on difficult-to-document methods of "visual inspection."

[19] Following Zhang and Levinson (2004), we restrict the pre-queue measurement period to be shorter than the post-queue measurement period.

we exclude queues that do not start between 2:15 pm and 7:00 pm. Queues during other hours of the day at these sites are more likely to be the result of construction, accidents, and other relatively unusual factors.

Fig. 6 plots median vehicle speeds by event time. To construct this figure we estimated an event study regression as in Eq. (1), but we specified our dependent variable as speed rather than flow. At all three sites speeds decline quickly over a relatively short time horizon near the onset of the queue. For example, at Site 1, median speeds exceed 55 miles-per-hour until shortly before queue formation, and then decrease sharply to below 20 miles-per-hour. Results are similar when we use means rather than medians (see Appendix Fig. A2).

The sharp speed decrease observed at all three sites is important because it suggests that the queue formation is a reasonably discrete event and that our results will not be unduly sensitive to the 30 miles-per-hour threshold. Indeed, later in the paper we assess the sensitivity of our results to alternative thresholds for defining a queue, and whether we use 25 miles-per-hour, 30 miles-per-hour, or 35-miles-per-hour, the results are quite similar.

Sites 1 and 3 exhibit sustained speed decreases for the full 80 min following queue formation. Site 2, in contrast, exhibits speeds that recover approximately 40 to 50 min following queue formation, implying that the longest queue of the day at this location tends to last less than one hour. Appendix Fig. A3 plots queue presence by event time and confirms this interpretation — queues generally persist for at least 80 min at Sites 1 and 3, but often dissipate in less than 80 min at Site 2. These shorter duration queues pose no specific issue for the event study analyses that follow, but they do imply that average flows after queue formation will tend to fall below capacity over time at Site 2.

5. Main results

5.1. Visual evidence

Fig. 7 presents our main results. The figure plots coefficients from our event study regression, Eq. (1). The horizontal axis measures

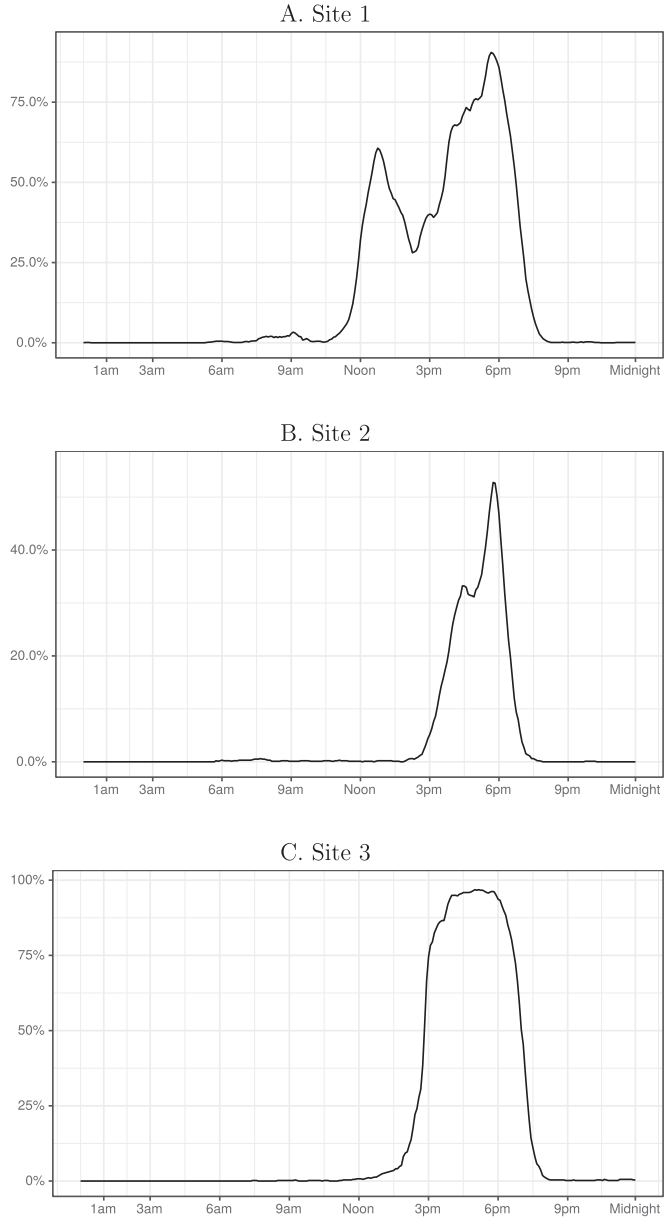

Fig. 4. Percentage of hours with a queue present. Notes: We exclude weekends and holidays. We define a queue as traffic moving under 30 miles-per-hour.

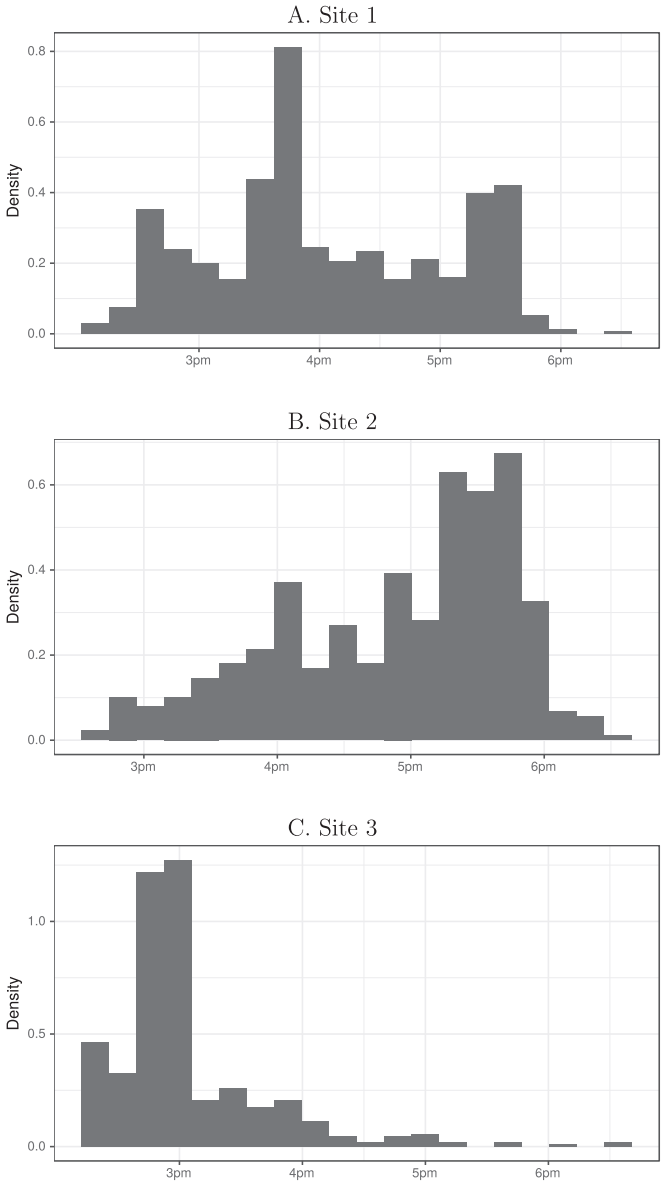

Fig. 5. Time-of-day that the queue begins each day, histogram. Notes: For each day, we select the longest continuous period of time with a queue, and then we define the start of the queue as the beginning of that period.

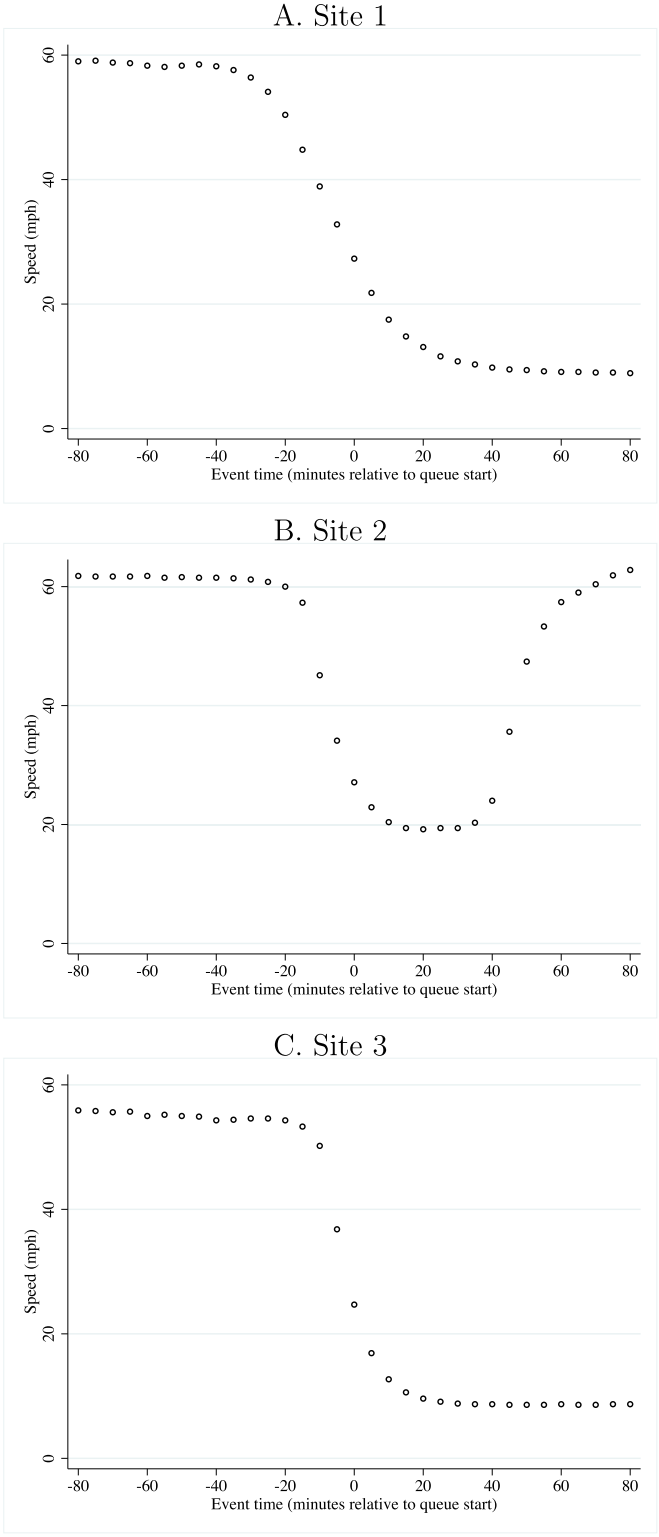

Fig. 6. Median vehicle speeds at queue onset. Notes: These event study figures plot median vehicle speeds in the 80 min before and after queue formation. Time is normalized so that the longest-duration afternoon queue begins at time zero on each day. Speed is measured at the nearest detector upstream of the bottleneck.

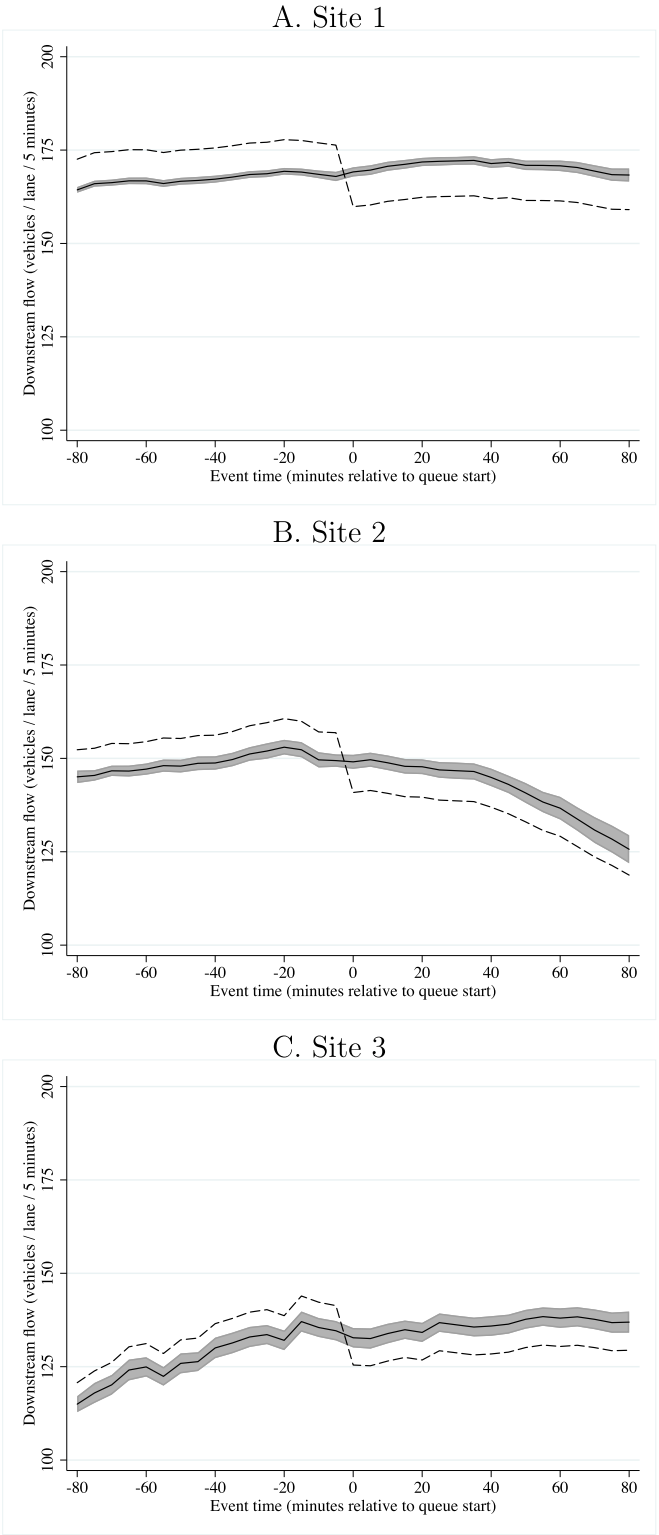

Fig. 7. Traffic flows by time of queue onset. Notes: These event study figures plot average vehicle flows in the 80 min before and after queue formation. Time is normalized so that the longest-duration afternoon queue begins at time zero on each day. The solid line plots average capacity, with the shaded area representing a 95% confidence interval, constructed using standard errors that are clustered by day-of-sample. The dashed line plots what average 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.

Table 2
Estimated change in highway capacity at queue formation.

Dependent Variable: Vehicles/5 min/lane	(1) 10-Minute Window	(2) 20-Minute Window	(3) 30-Minute Window	(4) 40-Minute Window
A. Site 1				
	1.23	1.17	1.24	1.40
	(0.36)	(0.35)	(0.37)	(0.37)
In-window mean	168.5	168.8	169.1	169.4
Number of days	706	706	706	706
B. Site 2				
	0.76	0.34	−1.68	−3.11
	(0.43)	(0.40)	(0.39)	(0.39)
In-window mean	142.0	142.9	144.2	145.0
Number of days	694	694	694	694
C. Site 3				
	−1.88	−2.42	−2.68	−1.30
	(1.09)	(0.86)	(0.72)	(0.66)
In-window mean	133.7	133.8	134.4	134.2
Number of days	247	247	247	247

Notes: This table reports twelve estimates of the change in highway capacity at queue formation, defined as the moment when average speed falls below 30 mph. 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), (3), and (4) expand the comparison to consider 20-, 30-, and 40-min symmetric windows (10, 15, and 20 min in each direction), respectively. Standard errors are clustered by day-of-sample.

event time. 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 near the moment the queue forms. Fig. 7 also includes 95% confidence intervals, and these intervals are narrow enough to rule out even modest changes in capacity. To illustrate this, we include 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.

Sites 1 and 3 demonstrate sustained flows near the observed maximum for a full 80 min following queue formation. At Site 2, average flows begin to fall below the observed maximum approximately 40 min after the queue forms. We do not interpret this decrease as capacity drop. Instead, as discussed earlier, queues at Site 2 tend to last less than 80 min, resulting in average flows that fall somewhat below capacity over time.

5.2. Baseline estimates

Table 2 reports estimates and standard errors that correspond with Fig. 7. As with Fig. 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 ($\hat{\beta}_{-1}$) and the first estimated β after queue formation ($\hat{\beta}_0$). Columns (2), (3), and (4) expand the comparison to consider 20-, 30-, and 40-min symmetric windows, respectively. In these columns we calculate the difference between the average estimated β coefficients before and after queue formation, in order to report the implied change in capacity per five minutes. For example, Column (2) takes the difference between $(\hat{\beta}_{-1} + \hat{\beta}_{-2})/2$ and $(\hat{\beta}_0 + \hat{\beta}_1)/2$. Positive (negative) estimates indicate an increase (decrease) in capacity.

Across study sites and specifications the estimates are tightly clustered around zero. Consistent with the visual evidence in Fig. 7, Table 2 reveals no evidence of a decrease in capacity when the queue forms. For example, for Site 1 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 1% of average capacity. Results are similar with alternative windows and for the other sites — there is a mix of positive and negative estimates, but all are negligible relative to average capacity. For all twelve estimates in Table 2 we can rule out a 5% capacity drop or larger with 99% confidence. If we average the estimates in each column across the three sites, we can reject an average capacity drop across sites of 1% or larger with 99% confidence in Columns (1) and (2) and 95% confidence in Columns (3) and (4).

There are tradeoffs to using a shorter or longer estimation window (e.g. 10, 20, 30, or 40 min). In a canonical event study design, a shorter estimation window minimizes the potential for bias due to secular time trends, but it comes at the cost of lower precision. In our context, the queue formation event is sudden but not instantaneous. Thus, in addition to the traditional bias-variance tradeoff, we also face the possibility that the queue formation event may not be fully captured in a 10-min window. It is therefore reassuring that our estimates across all

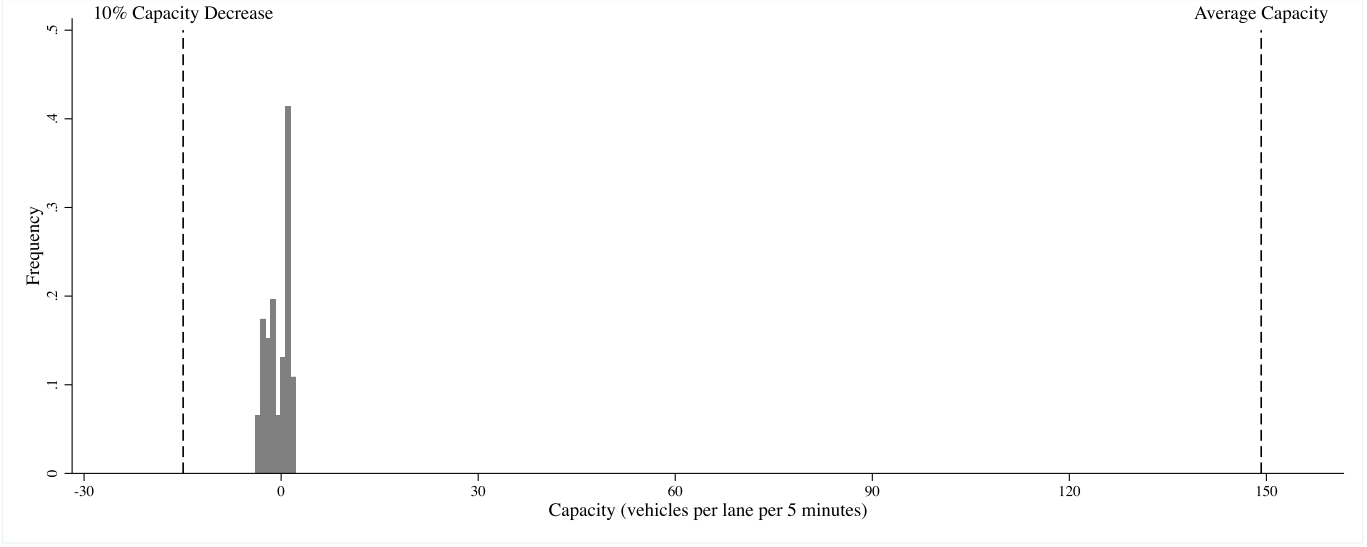

Fig. 8. Distribution of estimates. Notes: This figure plots the distribution of all 60 coefficient estimates from Tables 2, A2, A3, A4, and A5. The righthand vertical line corresponds to the average observed capacity across all sites and aforementioned tables. The lefthand vertical line corresponds to a hypothetical 10% decrease in capacity.

columns in Table 2 are close to zero in magnitude, implying that our conclusions are robust to estimation-window width.

5.3. Robustness checks

To further establish the robustness of our results, we consider a trimmed sample in which average speeds drop by over 20 miles-per-hour in less than 20 min. By construction this trimmed sample drops days on which the queue forms more gradually. Appendix Fig. A4 plots the median speed by event time for each site, after applying this trimming rule. For Site 1 (SR-24) this constraint makes the drop in speeds even sharper, while for the other two sites the constraint is essentially non-binding. Appendix Fig. A5 and Appendix Table A2 present analogous results for traffic flows when trimming the sample to days on which average speeds drop by over 20 miles-per-hour in less than 20 min. There is again no evidence of a capacity drop at queue onset.

Appendix Tables A3 and A4 report regression estimates using alternative thresholds to define a queue. Whereas our baseline estimates define a queue to be present whenever traffic is moving at under 30 miles-per-hour, these alternative specifications adopt thresholds of 25 miles-per-hour and 35 miles-per-hour. With the more restrictive threshold (25 miles-per-hour), 11 of 12 point estimates move closer to zero. With the less restrictive threshold (35 miles-per-hour) half the point

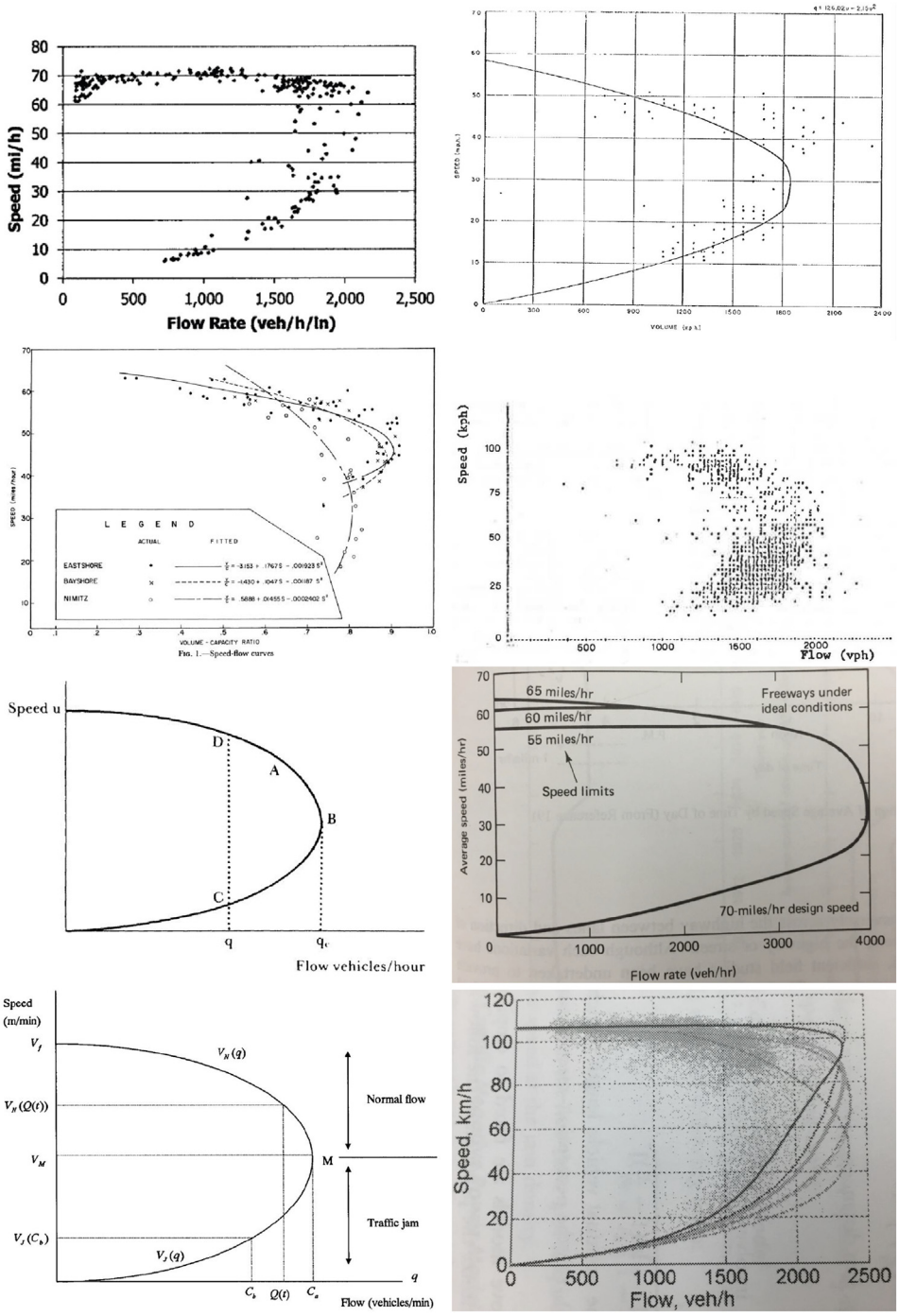

Fig. 9. Speed-flow curves. Sources: Transportation Research Board (2016), Drake and Schofer (1966), Keeler and Small (1977), Allen et al. (1985), Newbery (1989), May (1990), Mun (1999), and Ni (2015).

estimates move further from zero, and half are unchanged or closer to zero.

Appendix Table A5 reports results from alternative event study analyses in which we estimate the specification used in Table 2 with 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. Six of the twelve estimates are positive, and in all cases we can reject a 5% capacity drop or larger.

5.4. Effect size magnitudes

Across specifications, the event study analyses demonstrate no evidence of a decrease in highway capacity upon queue formation. To put these results in context, Fig. 8 plots a histogram of all of our estimates. We include all coefficient estimates from Table 2, as well as from alternative analyses in Appendix Tables A2, A3, A4 and A5 (60 coefficients in total). These estimates summarize the results from our test across three sites and a rich variety of specifications.

All of our estimates are clustered tightly around zero. To illustrate this fact we include in Fig. 8 two vertical lines. The righthand vertical line corresponds to the average observed capacity across all sites and tables — 149 vehicles per lane per five minutes. The lefthand vertical line corresponds to a hypothetical 10% decrease in capacity (14.9 vehicles per lane per five minutes). Even the most negative of our 60 estimates fall well short of this 10% threshold, and the vast majority of estimates are either positive or represent less than a 2% decrease in average capacity.

6. Discussion

6.1. 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 which drivers face a tradeoff between time delays and schedule inflexibility and optimize their departure times accordingly.

Economists have long recognized that traffic congestion represents a negative externality (Pigou, 1920; Vickrey, 1963, 1969). 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 at least in the context of isolated highway bottlenecks, this externality does not appear to be exacerbated by an additional decrease in capacity. Driving reduces average speeds, but we find no evidence at our three sites of a drop in capacity at the onset of queueing. 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. Arnott et al., 1993). 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 at the optimum.

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, 2018, forthcoming). For example, Hall (forthcoming) 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.

Finally, it is worth emphasizing that even without hypercongestion, the standard negative externality from traffic congestion can be very large. For example, the queues in our three study sites routinely reach one hour or more in length. Thus when a motorist decides to drive these routes during peak periods they impose a delay on other drivers equal to a total of up to one hour or more. Our paper is not focused on this standard negative externality, but we bring this up because our evidence on the lack of evidence of one form of hypercongestion should not be interpreted as suggesting that this standard externality does not exist or is small in magnitude.

7. Speed-flow curves

Before concluding, we perform one additional graphical analysis. In the transportation and economics literatures it is common to plot "speed-flow" curves depicting the locus of speed-flow observations over some time period at a particular location. Fig. 9 shows eight examples.[20] In all cases, the horizontal axis measures traffic flow and the vertical axis measures speed.

The upper part of the speed-flow curve typically exhibits a negative correlation between speed and flow. Speeds are high at low flow levels and then decrease at higher flow levels. The lower part of the curve is more surprising, however — this part exhibits a *positive* correlation between speed and flow. Particularly striking are the observations with both very low speeds and very low flows. The *Highway Capacity Manual* explains that this lower 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.

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. More recent economic analyses, however, have argued conceptually that this relationship should not 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."[21] Similarly, Lindsey and Verhoef (2008) summarizes an "emerging view" that these low-

[20] From the transportation literature, Drake and Schofer (1966); Allen et al. (1985); May (1990); Ni (2015); Transportation Research Board (2016). From the economics literature, Keeler and Small (1977); Newbery (1989); Mun (1999). Also see Russo et al. (2019).

[21] 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).

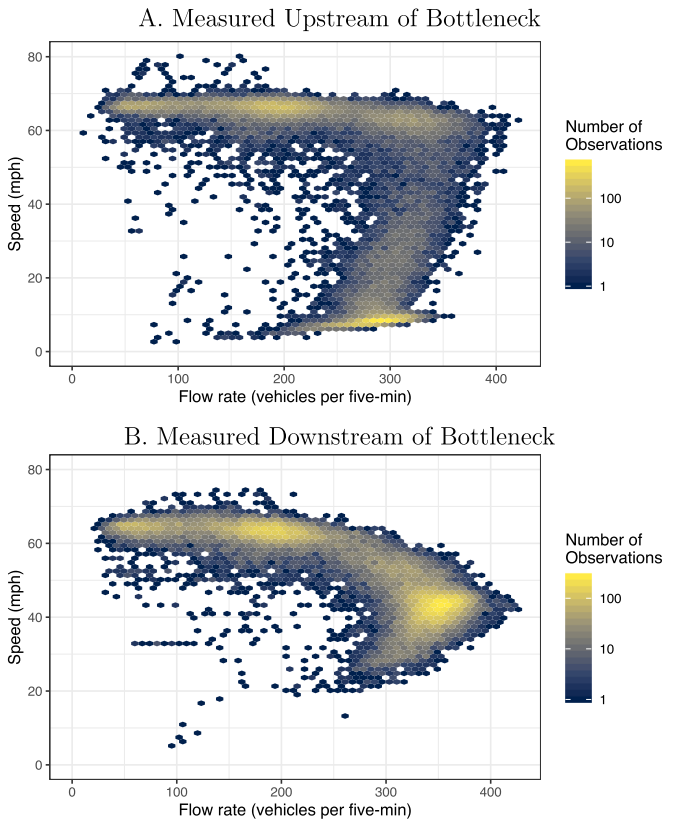

Fig. 10. Observed speed–flow curves. Notes: This figure plots traffic flows and vehicle speeds from Site 1. The unit of observation is a speed-flow pair averaged over five minutes. The first panel plots observations from the last upstream detector before the bottleneck, and the second panel plots observations from the first downstream detector after the bottleneck. We plot data for weekdays between 1 pm and 11:55 pm (a period when the lane reduction is in effect) and restrict the sample to be identical in both panels. Colors represent the number of observations in each cell, as indicated in the legend.

speed, low-flow observations occur "in queues upstream of a bottleneck". (p. 421).

We provide empirical support for the view that low-speed, low-flow observations represent queuing and have no direct implications regarding capacity. Fig. 10 presents two speed-flow curves for Site 1. This site works well for constructing speed-flow curves because the downstream loop detector is located well past the bottleneck and there is no intervening merging traffic.[22]

The top panel of Fig. 10 uses data from the detector that is just upstream of the bottleneck. With hundreds of days of data measured at 5-min intervals, 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 that appear in Fig. 9. There is a large mass of observations at 60 miles-per-hour or faster, and speeds decrease modestly with flow rates along the top part of the speed-flow curve. But then, as is typical in speed-flow curves, there are also large numbers of low-speed, low-flow observations which make the curve bend backward. Particularly striking are the observations with both very low speeds and very low flows. For example, there is a considerable mass of observations with speeds below 10 miles-per-hour and flow rates below 250 vehicles per five minutes.

The bottom panel of Fig. 10 is identical to the top panel, except it is constructed using data from the downstream detector. The pattern in this second panel is quite different. In particular, there are very few observations with below 250 vehicles per five minutes and speeds below 40 miles-per-hour. The divergence between the two panels suggests a simple explanation: the low-flow, low-speed observations represent traffic waiting in the queue. At the downstream detector, where queues rarely form, there are virtually no low-flow, low-speed observations, but at the upstream detector they are numerous. Indeed, the entire region which the *Highway Capacity Manual* refers to as "flow breakdown" or "oversaturated flow" essentially does not exist in the bottom panel.

This simple comparison provides a simple illustration of why speed-flow curves should not be interpreted as causal relationships. Low-flow, low-speed observations measured upstream of a bottleneck do not provide evidence for or against capacity drop or hypercongestion. These observations occur in the queue so do not provide information about the rate of flow through the bottleneck.

8. Conclusion

The concept of hypercongestion has influenced transportation economics models for over five decades. Our paper proposes an empirical test of hypercongestion at highway bottlenecks. Our test is designed for highway bottlenecks with a single, well-defined bottleneck — not dense urban areas or locations with multiple bottlenecks and queue spillovers. Consequently, our results speak only to one of the two forms of hypercongestion that have been discussed in the literature.

Our test is novel in the literature but uses standard event study methodologies that have been widely used in other contexts. We apply our test to high-quality data from three highway bottlenecks in California. We document significant speed decreases at all three sites during weekday afternoons. However, we find no evidence of a decrease in traffic flows at the onset of queue formation. Results are similar across all three sites and a range of alternative specifications, with no evidence of a drop in capacity.

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 forms.

Our findings imply that marginal damages at highway bottlenecks are much lower than implied by supply curves exhibiting hypercongestion. Nevertheless, congestion taxes should not be zero. To the contrary, 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpubeco.2020.104197.

References

Akbar, Prottoy, Duranton, Gilles, 2017. Measuring the Cost of Congestion in Highly Congested City: Bogotá. University of Pennsylvania Working Paper.
Allen, B.L., Hall, F.L., Gunter, M.A., 1985. Another look at identifying speed-flow relationships on freeways. Transp. Res. Rec. 1005, 54–64.

[22] In contrast, Site 2 is a merge with another major highway, so downstream traffic includes vehicles from both highways, and at Site 3 the downstream loop detector is quite close to the bottleneck, so vehicles are still accelerating as they pass the detector.

Arnott, Richard, 2013. A bathtub model of downtown traffic congestion. J. Urban Econ. 76, 110–121.
Arnott, Richard, De Palma, Andre, Lindsey, Robin, 1990. Economics of a bottleneck. J. Urban Econ. 27 (1), 111–130.
Arnott, Richard , Andre De Palma, and Robin Lindsey, "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 1993, pp. 161–179.
Arnott, Richard , Andre 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., 1990. Flow processes at a freeway bottlenecks. Transp. Res. Rec. (1287).
Banks, James H, "Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering?," Transp. Res. Rec., 1991, (1320).
Baum-Snow, Nathaniel, 2007. Did highways cause suburbanization? Q. J. Econ. 122 (2), 775–805.
Bertini, Robert L and Monica T Leal, "Empirical Study of Traffic Features at a Freeway Lane Drop," J. Transp. Eng., 2005, 131 (6), 397–407.
Bertini, Robert, Malik, Shazia, 2004. Observed dynamic traffic features on freeway section with merges and diverges. Transp. Res. Rec. 1867, 25–35.
Cassidy, Michael J., Bertini, Robert L., 1999. Some traffic features at freeway bottlenecks. Transp. Res. B 33 (1), 25–42.
Cassidy, Michael J., Rudjanakanoknad, Jittichai, 2005. Increasing the capacity of an isolated merge by metering its on-ramp. Transp. Res. B 39 (10), 896–913.
Chin, Hong C., May, Adolf D., 1991. Examination of the speed-flow relationship at the Caldecott tunnel. Transp. Res. Rec. 1320, 75–82.
Chung, K., Cassidy, M., 2002. "Testing Daganzo's Behavioral Theory Multi-lane Freeway TrafficCalifornia Partners for Advanced Transit and Highways (PATH).
Chung, Koohong, Rudjanakanoknad, Jittichai, Cassidy, Michael J., 2007. Relation between traffic density and capacity drop at three freeway bottlenecks. Transp. Res. B 41 (1), 82–95.
Couture, Victor, Duranton, Gilles, Turner, Matthew A., 2018. Speed. Rev. Econ. Stat. 100 (4), 725–739.
Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J Notowidigdo, "The Economic Consequences of Hospital Admissions," Am. Econ. Rev., 2018, 108 (2), 308–52.
Drake, J.L., Schofer, Joseph L., 1966. A statistical analysis of speed-density hypotheses. Highw. Res. Rec. 154, 53–87.
Duggan, Mark, Garthwaite, Craig, Goyal, Aparajita, 2016. The market impacts of pharmaceutical product patents in developing countries: evidence from India. Am. Econ. Rev. 106 (1), 99–135.
Duranton, Gilles, Turner, Matthew A., 2011. The fundamental law of road congestion: evidence from US cities. Am. Econ. Rev. 101 (6), 2616–2652.
Fosgerau, Mogens, Small, Kenneth A., 2013. Hypercongestion in downtown Metropolis. J. Urban Econ. 76, 122–134.
Freyaldenhoven, Simon, Hansen, Christian, Shapiro, Jesse M., 2019. Pre-event trends in the panel event-study design. Am. Econ. Rev. 109 (9), 3307–3338.
Guan, Yu, Jishuang Zhu, Ning Zhang, and Xiaobao Yang, "Traffic flow characteristics of bottleneck segment with ramps on urban expressway," in "International Conference on Transportation Engineering 2009" 2009, pp. 1679–1684.
Hall, Jonathan D., 2018. Pareto improvements from Lexus lanes: the effects of pricing a portion of the lanes on congested highways. J. Public Econ. 158, 113–125.
Hall, Jonathan D., "Can Tolling Help Everyone? Estimating the Aggregate and Distributional Consequences of Congestion Pricing," *Journal of the European Economic Association*, (forthcoming) 2020.
Hall, Fred L and Kwaku Agyemang-Duah, "Freeway Capacity Drop and the Definition of Capacity," *Transportation Research Record*, 1991, (1320).
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.
Hurdle, V.F., Datta, P.K., 1983. Speeds and flows on an urban freeway: some measurements and a hypothesis. Transp. Res. Rec. 905, 127–137.
Jin, Wen-Long, Gan, Qi-Jian, Lebacque, Jean-Patrick, 2015. A kinematic wave theory of capacity drop. Transp. Res. B Methodol. 81, 316–329.
Johnson, M. Bruce, 1964. On the economics of road congestion. Econometrica 32 (1,2), 137.
Keeler, Theodore E and Kenneth A Small, "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways," J. Polit. Econ., 1977, 85 (1), 1–25.
Kreindler, Gabriel, 2018. The Welfare Effect of Road Congestion Pricing: Experimental Evidence and Equilibrium Implications. MIT Working paper.
Lamotte, Raphaël, Andre De Palma, and Nikolas Geroliminis, "On the Use of Reservation-Based Autonomous Vehicles for Demand Management," Transp. Res. B Methodol., 2017, 99, 205–227.
Leclercq, Ludovic, Victor L Knoop, Florian Marczak, and Serge P Hoogendoorn, "Capacity Drops at Merges: New Analytical Investigations," Transportation Research Part C: Emerging Technologies, 2016, 62, 171–181.
Lindsey, C Robin and Erik T Verhoef, "Congestion Modeling," in "Handbook of Transportation Modelling" Elsevier Science 2008.
Lomax, Tim, Schrank, David, Eisele, Bill, 2018. Congestion Data For Your City — Urban Mobility Information.
May, Adolf D, *Traffic Flow Fundamentals*, Prentice Hall, 1990.
Mun, Se-il, 1999. Peak-load pricing of a bottleneck with traffic jam. J. Urban Econ. 46 (3), 323–349.
Newbery, David M, "Cost Recovery from Optimally Designed Roads," Economica, 1989, 56 (222), 165–185.
Ni, Daiheng, 2015. Traffic Flow Theory: Characteristics, Experimental Methods, and Numerical Techniques. Butterworth-Heinemann, Elsevier.
Oh, Simon, Yeo, Hwasoo, 2012. Estimation of capacity drop in highway merging sections. Transp. Res. Rec. 2286, 111–121.
Persaud, Bhagwant, Yagar, Sam, Brownlee, Russel, 1998. Exploration of the breakdown phenomenon in freeway traffic. Transp. Res. Rec. 1634, 64–69.
Pigou, Arthur Cecil, 1920. The Economics of Welfare. Macmillan, London.
Russo, Antonio, Adler, Martin, Liberini, Federica, van Ommeren, Jos N., 2019. Welfare Losses of Road Congestion. Working Paper.
Small, Kenneth A, "The Scheduling of Consumer Activities: Work Trips," Am. Econ. Rev., 1982, 72 (3), 467–479.
Small, Kenneth A, "The Bottleneck Model: An Assessment and Interpretation," Econ. Transp., 2015, 4 (1-2), 110–117.
Small, Kenneth and Xuehao Chu, "Hypercongestion," Journal of Transport Economics and Policy, 2003, 37 (3), 319–352.
Srivastava, Anupam, Geroliminis, Nikolas, 2013. Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model. Transportation Research Part C: Emerging Technologies 30, 161–177.
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 J. Phys., 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 J. Phys., 2013, 15 (10), 103034.
Transportation Research Board, "Highway Capacity Manual 6th Edition: A Guide for Multimodal Mobility Analysis," 2016.
Vickrey, William S, "Pricing in Urban and Suburban Transport," Am. Econ. Rev., 1963, 53 (2), 452–465.
Vickrey, William S, "Congestion Theory and Transport Investment," Am. Econ. Rev., 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, Jun, Avralt-Od Purevjav, and Shanjun Li, "The Marginal Cost of Traffic Congestion and Road Pricing: Evidence from a Natural Experiment in Beijing," *American Economic Journal: Economic Policy*, (forthcoming) 2020.
Yuan, Kai, Victor L Knoop, and Serge P Hoogendoorn, "Capacity Drop: Relationship Between Speed in Congestion and the Queue Discharge Rate," Transp. Res. Rec., 2015, 2491 (1), 72–80.
Zhang, Lei, Levinson, David, 2004. Some properties of flows at freeway bottlenecks. Transp. Res. Rec. 1883, 122–131.