

Congestion, excess demand, and effective capacity in California freeways*

Zhanfeng Jia, Pravin Varaiya, Chao Chen, Karl Petty, and Alex Skabardonis
PeMS Development Group
University of California, Berkeley

December 9, 2000 (rev)

Abstract

The paper makes four assertions, supported by an extensive empirical study of freeways in Los Angeles and Orange County. First, maximum throughput occurs at the free flow speed of 60 mph, and not between 35 and 45 mph, as is often assumed. So congestion must be measured as the additional vehicle-hours of delay traveling below 60 mph.

Second, the maximum throughput over a link—its effective capacity—depends on how a link is connected to other links and the pattern of traffic, as well as its physical characteristics. A challenge to traffic theory is to determine the maximum throughput of a link, given the network topology and traffic pattern.

Third, the congestion delay can be divided into (1) the portion that can be eliminated by ramp metering, and (2) the delay due to excess demand. There is a systematic procedure to calculate these two components of delay for recurrent congestion, using loop-detector data. The procedure is illustrated for a 6-mile stretch of I-405N in Orange County on June 1, 1998, 5.00-10.00 am.

Fourth, recurrent congestion evolves over three phases: increased demand is first met at free flow speeds, until the demand exceeds maximum throughput and congestion starts; both flow and speed then decrease and occupancy increases; only after demand drops well below maximum throughput does occupancy decrease and speed increase until free flow is reestablished. So the objective of ramp metering must be to maintain free flow and maximum throughput. It is a challenge to design such a ramp metering algorithm.

*The work reported here is a joint effort of the PeMS Development Group. Tom Choe, Joe Hecker, Robert Kopp, Tom West, and John Wolf of Caltrans; Tarek Hatata, Elizabeth Stoltzfus, and Chris Williges of Booz, Allen and Hamilton; Patrick Conroy of California PATH; and Professors Ben Coifman, Markos Papageorgiou and Michael Zhang helped us with their encouragement, criticism, and suggestions. The PeMS project is supported by Caltrans, National Science Foundation Grant CMS-0085739, and EPRI/DoD Complex Interactive Networks Initiative under Contract WO8333-04. The authors alone are responsible for the opinions expressed here.

1 Introduction

There is interest within the California Department of Transportation or Caltrans in measuring free-way congestion, setting realistic congestion-reduction targets, and making appropriate investments to meet those targets. But opinions differ over how to measure congestion; how much congestion can be eliminated by metering and how much of it is due to excess demand, requiring demand management for its reduction; and the relative magnitudes of recurrent vs non-recurrent congestion. The reason for these differences is simple: in the absence of reliable empirical knowledge of congestion and its causes, people holding different opinions won't change them. This paper summarizes evidence to resolve some of these differences. It is a report of the PeMS (Performance Measurement System) project—a large-scale study to measure the performance of California freeways [2].

Congestion measures compare the actual travel time to some standard. There are two defensible standards: one is travel time under free flow conditions, the other is travel time under maximum throughput. We demonstrate that in California, the two standards coincide: maximum throughput occurs at the free flow speed of 60 mph. So the standard proposed here is to measure congestion as the additional vehicles hours spent traveling below 60 mph. Caltrans today declares a link congested if its speed is below 35 mph for at least 15 min. This practice cannot be justified and should be replaced by 60 mph.

We call the maximum observed sustainable throughput in a link its **effective capacity**. It is generally quite different from the Highway Capacity Manual's definition of capacity which is a function of the link's physical characteristics. Effective capacity depends on how a link is interconnected with other links, the pattern of demand, and its physical characteristics. Effective capacity is an empirical notion—it is the maximum observed throughput over a link during recurrent congestion. The fact that this maximum throughput is remarkably stable, despite daily fluctuations, suggests that the notion reflects some more stable phenomenon. It is a challenge to theoretically explain effective capacity.

Congestion can be reduced or eliminated by proper ramp-metering and by demand diversion. Put inversely: imperfect metering and excess demand cause congestion. The PeMS procedure provides a theoretically sound way to divide congestion between these two causes. Moreover, numerical calculation of the total congestion and its two shares is straightforward. So the procedure can help to set targets for, and to monitor the effectiveness of, both ramp-metering and demand-diversion policies.

Recurrent congestion fluctuates every day. So it is difficult empirically to calculate the additional delay caused by an incident during recurrent congestion. The procedure described here allows the calculation of the probability distributions of recurrent congestion, with and without incidents, provided incident data are available. In principle, the two distributions would tell us how much congestion is non-recurrent.

PeMS collects 30-sec, loop-detector data from California freeways. These data are analyzed in real-time and stored in the PeMS database. Many applications of interest to Caltrans management, engineers, planners, and the public are accessible over the World Wide Web at <http://transacct.eecs.berkeley.edu>. Other applications, like the one reported here, require direct access to the PeMS database.

2 Congestion delay and its components

Freeway congestion typically starts in a “chokepoint” link when flow into the link exceeds its effective capacity. Density builds up in the chokepoint, and congestion infects upstream links, one by one. Recovery typically proceeds in the reverse direction. Demand first drops below effective capacity in the most upstream congested links, and they become uncongested. The reduced demand propagates downstream, relieving links one by one.

Figure 1 shows the contour plots of a section of I-405N, extending from postmile 0 to 7, during the time interval 5.00-10.00 am, for 22 weekdays in June, 1998. For each contour plot, the x-axis is time, the y-axis is distance (vehicles travel from bottom to top). Darker points correspond to lower speeds.

There is free flow at 5.00 am, with vehicles traveling at 60 mph. The morning commute congestion starts shortly before 7.00 am near postmile 5. The contagion spreads upstream. At the depth of the congestion speed is below 15 mph. Demand eventually falls sufficiently to bring relief to the most upstream links of the congested region, and relief propagates downstream. Most days, free flow is restored by 9.30 am. Although there is congestion every morning, the contour plots also reveal daily fluctuations.

We analyze the congestion on the morning of June 1, 1998. Figure 2 shows three graphs. The top graph is the time (in vehicle-hours-traveled or VHT) that vehicles actually spent on this section in each 5-minute time interval between 5.00 and 10.00 am. (That number is multiplied by 12 so the units are VHT per hour instead of VHT per five minutes.) The middle graph is the VHT they would have spent in each 5-minute interval under ideal metering that maintains maximum flow at free flow speed of 60 mph on the freeway. (Ideal metering is described later.) Under ideal metering, some vehicles would spend time at the ramps, and the middle graph includes that time. The bottom graph is the time that vehicles would spend traveling in the study section under free flow conditions (60 mph). So the bottom graph is the same as the middle graph, excluding time spent at the ramps.

The area under the top graph is the actual total VHT spent on this section of I-405N on June 1, 1998 between 5.00-10.00 am. The area under the bottom graph is the total time they would have spent traveling at 60 mph. The difference between the two areas is the total **congestion delay**, $Delay_{tot}$. Define

$$\begin{aligned} Delay_{tot} &= \text{Area under top graph} - \text{Area under bottom graph}, \\ Delay_{met} &= \text{Area under top graph} - \text{Area under middle graph}, \\ Delay_{dem} &= \text{Area under middle graph} - \text{Area under bottom graph}. \end{aligned}$$

The total delay can be divided into two parts,

$$Delay_{tot} = Delay_{met} + Delay_{dem}. \quad (1)$$

$Delay_{tot}$ is the congestion delay—the extra time spent traveling below 60 mph. $Delay_{met}$ is the delay that can be eliminated by ideal metering. $Delay_{dem}$ is the delay due to excess demand (demand larger than effective capacity); it can only be avoided by shifting demand.

$Delay_{tot}$	704	534	808	353	340	266	595	252	1551	324
$Delay_{met}$	549	446	656	323	209	247	550	138	1266	87

Table 1: $Delay_{tot}$ and the delay remaining after ideal metering $Delay_{met}$ for 10 days in June 1998.

An eyeball estimate from figure 2 gives

$$Delay_{tot} = 500 \text{ VHT}, \quad Delay_{met} = 350 \text{ VHT}, \quad Delay_{dem} = 150 \text{ VHT}.$$

So on June 1, 1998, out of the total congestion delay of 500 VHT, 350 VHT could, in principle, be eliminated by ramp metering. The remaining 150 VHT of delay is due to demand in excess of the capacity of this freeway section. It can only be avoided by diverting demand, perhaps by telling motorists how much time they would spend parked at the ramps.

Table 1 gives estimates of these delays for 10 different days in June 1998. On average, about two-thirds of the total congestion delay could be eliminated by ideal metering. Of course, in other freeway sections, the delay shares may be different.

The remainder of the paper explains the underlying concepts and the procedure used to obtain these delays using PeMS loop detector data.

3 Effective capacity

The PeMS procedure relies on the concept of the **effective capacity** of a link. We develop this concept in the context of the I-405N study section during the congestion episode on June 1, 1998, 5.00-10.00 am. The study section is depicted in figure 3. The figure shows the location of the detectors on the mainline and on the ramps.

Each link has four or five lanes and one HoV lane. There are 13 links and eight on- and off-ramps. The first on-ramp is virtual, representing metering of traffic upstream of the study section, presumably implemented by ramps there. A **link** is a portion of freeway associated with the loops at one location and extending half way to the next upstream and downstream loops, as in figure 4. A link may contain at most one on- or off-ramp. The links in the study section are named ML1, \dots , ML13 (ML denotes mainline detectors).

The top left graph in figures 6–18 plots 5-minute averages of the actual flow in vehicles per hour (VPH) vs occupancy (percent) for the 13 links ML1–ML13. (Ignore for now the other three plots in the figures.) Initially vehicles travel at 60 mph.¹ Flow increases until it reaches a maximum value and congestion starts. Speed and flow now drop while occupancy increases. Eventually demand drops and speed gradually increases until the free flow regime is recovered.

In all cases the plots look like that shown in figure 5. We have examined flow vs occupancy

¹The slopes of the initial straightline portion in the plots all correspond to a speed of 60 mph. In general, the slope of the line from the origin to any point in the flow-occupancy plane is proportional to the speed at that point.

in 4,000 links in Los Angeles county. They all behave like in the figure. We conclude that in California, maximum throughput occurs at the freeflow speed of 60 mph. Hence

Congestion delay should be measured as the additional vehicles hours traveled driving below 60 mph.

We define the **effective capacity of a link** to be the maximum sustainable flow in VPH that is reached in that link. PeMS data are used to calculate the effective capacity of the 13 links in the study section during the congestion episode. From the flow vs occupancy plots we find that the maximum flow for ML1 and hence its capacity is 9,300 VPH, the capacity of ML2 is 10,000 VPH, the capacity of ML3 is 10,000 VPH, and so on. In this way we obtain the effective capacity C_k for every link k .

Effective capacity is thus an empirical concept and its measurement will vary from day to day. But if the concept is sound, the effective capacity numbers should be close to each other during *recurrent* congestion episodes.² And that is indeed the case. Figure 27 plots the PeMS capacity numbers for the 13 links in the study section for five days in June 1998 during the recurrent congestion of the morning commute.

The remarkable agreement in the calculations for different days for each of the 13 links lends credence to the proposition that the effective capacity is a stable characteristic of traffic behavior. But instead of being constant, it is more in accord with the data to regard effective capacity as a *stochastic* quantity with a narrow range of daily variability.³

Observe that the capacities of the 13 links according to their physical characteristics are virtually the same (Figure 3), but their effective capacities vary by as much as 50 percent, so predictions based on physical capacity can mislead. It is a challenge to traffic theory to explain the stability of the empirical measurement of effective capacity and to determine the effective capacity of a link from the network topology and traffic pattern.

4 Ideal metering and excess demand

The notion of effective capacity suggests the following hypothesis about traffic behavior:

If a metering policy keeps flow below its effective capacity in every link throughout a congestion episode, the speed will be maintained at 60 mph, i.e. congestion will disappear. A consequence of the metering is that vehicles will be stopped at the ramps for some time.

We call this the *Ideal Metering Principle* (IMP). Of course we don't know if IMP holds in practice, but the evidence lends it plausibility. The ideal metering policy is based on this principle. For on-ramp r , let $d(r)$ be the link downstream of r (in figure 3 this is the link containing on-ramp r) and let

²Of course, if there are incidents or lane closures, the effective capacity can be quite different.

³The variability may be due to microscopic shifts in driver behavior, traffic patterns, weather, etc. This belief is supported by an observation based on the figure—the capacities of all links move together, i.e. on some days they are all slightly larger or all slightly smaller. There is likely to be the same underlying cause.

$u(r)$ be the link upstream of r . Then in any period t , ramp r should be metered at rate $On_r(t)$ equal to the downstream link capacity $C_{d(r)}$ minus the flow on the upstream link $u(r)$. If traffic behavior satisfies IMP, then under ideal metering, flow on all links will not exceed effective capacity and traffic will move at 60 mph.

Queues may build up at ramps under ideal metering. We define **excess demand delay** as the queuing delay under ideal metering. This definition is appropriate because any attempt to reduce it by increasing the metering rate will lead to congestion and an increase in the total delay. Excess demand delay can only be reduced by shifting demand over time or space, or to other modes. We now estimate this delay.

We use the following data from PeMS for every 5-minute interval t from 5.00 to 10.00 am:

$$\begin{aligned} In_r(t) &= \text{inflow into on-ramp } r \text{ in } t, \\ Out_s(t) &= \text{outflow into off-ramp } s \text{ in } t, \\ Vol_k(t) &= \text{flow on link ML}k \text{ of study section in } t. \end{aligned}$$

All these quantities are in VPH. The first two quantities characterize the exogenous demand during the congestion episode: how many vehicles enter each on-ramp r and leave from each off-ramp s . Included in these are the virtual on-ramp upstream and off-ramp downstream of the study section.

*We assume that the exogenous demand is unchanged by the metering policy.*⁴

The link flows will of course be changed by metering and we denote them by a superscript, e.g., $Vol_k^{met}(t)$ is the flow in link ML k in t . These flows are easy to calculate:⁵

$$\begin{aligned} Vol_k^{met}(t) &= \text{Sum of metered flows } In_r^{met}(t) \text{ from all on-ramps } r \text{ upstream of ML}k \\ &\quad - \text{Sum of outflows } Out_s(t) \text{ into all off-ramps } s \text{ upstream of ML}k. \end{aligned} \quad (2)$$

The outflows $Out_s(t)$ are part of the data.⁶ The metered flows $In_r^{met}(t)$ are determined below in (5).

Next we calculate the queue build-up at each ramp. Consider on-ramp r , including the virtual ramp upstream of ML1. At time t this ramp will have a queue of $q_r^{met}(t)$ vehicles given by

$$q_r^{met}(t+1) = [q_r^{met}(t) + In_r(t) - On_r(t)]^+, \quad t = 0, 1, \dots \quad (3)$$

Here,

$$On_r(t) = C_{d(r)} - Vol_{u(r)}^{met}(t) \quad (4)$$

⁴Of course demand will increase in response to reduced travel times. So this assumption is just part of the ‘‘thought experiment’’ for calculating excess demand. The calculation could, in principle, be extended to incorporate a demand-response model.

⁵Equation (2) takes this simple form because at 60 mph a vehicle traverses the entire section within one 5-minute period. If the study section was, say, 20 miles long, so that the free flow travel time would take four 5-minute periods, the outflows in t would depend on the inflows at $t, t-1, t-2, t-3, t-4$ and the equation would be more complex.

⁶As Markos Papageorgiou observed, under ideal metering, the outflows will be larger than observed values, further reducing the delay. This reduction is ignored here.

is the ideal metering rate of on-ramp r in VPH, $In_r(t)$ is the demand in VPH at this ramp from PeMS data, and the notation $[x]^+ = \max\{x, 0\}$ guarantees that queues can't be negative. The boundary condition of (3) is $q_r^{met}(0) = 0$, since initially (at 5.00 am) there is no congestion and hence no queue.

Under the metering policy, the metered inflow from ramp r is given by

$$In_r^{met}(t) = \begin{cases} On_r(t), & \text{if } q_r(t) > 0 \\ In_r(t), & \text{if } q_r(t) = 0 \end{cases} . \quad (5)$$

That is, so long as there is a queue, inflow is at the metered rate $On_r(t)$; if there is no queue, inflow is at the measured demand $In_r(t)$.

Equations (2)–(5) determine what happens under ideal metering. Traffic in the freeway section under study moves at 60 mph. Queues build up when there is excess demand, and then dissipate according to (3). Knowing the queues at each t , we can calculate the waiting time at each ramp.

The top right plots in Figures 6–18 show both $Vol_k(t)$ (dotted line) and the flow after metering $Vol_k^{met}(t)$ (solid line). The metered flow on each link is always restricted to be below effective capacity—the maximum achieved flow in the adjacent plot on the left. When $Vol_k(t) > Vol_k^{met}(t)$, vehicles are held back at ramps upstream of k , and when $Vol_k(t) < Vol_k^{met}(t)$ those queues are dissipated.⁷

The bottom left plots in Figures 6–18 show the occupancy from PeMS data (dotted line) and the occupancy under the metering policy (solid line). Notice how metering “clips” the congestion-causing high-occupancy periods.

Since speed is a constant 60 mph under metering, occupancy is strictly proportional to flow $Vol_k^{met}(t)$ as is seen by comparing the bottom left and top right plots in Figures 6–18. Observe that occupancy is a much more sensitive measure of how close the flow is to effective capacity. So it is much better to *implement* the ideal metering policy (4) using measured occupancy downstream of the ramp rather than measured flow upstream of the ramp.

Figures 19–26 describe the queue behavior at the 8 on-ramps. In each case, the top left shows the demand $In_r(t)$ (dotted line) and the metered inflow $In_r^{met}(t)$ (solid line). Queues build up when $In_r(t) > In_r^{met}(t)$ and recede when the inequality is reversed. The top right is a plot of the queue $q_r^{met}(t)$ at each 5-minute interval t , calculated according to (3). The bottom right and left plots give the total waiting time (in vehicle-hours) and the per vehicle average waiting time (in minutes). These quantities are calculated assuming that the departures and arrivals of vehicles (top left) are uniformly distributed over each 5-minute interval.

At ramps 2, 4, 6, and 8, the maximum waiting time is between 4 and 6 minutes and the maximum queue sizes are also large (as many as 120 vehicles). The large queues occur when demand is at its peak and, in unmetered congestion, *everyone* would be spending an extra 5 minutes crawling at 13 mph on the freeway. In light of this, a maximum queue size of 120 at a ramp is remarkably small.

⁷This may seem paradoxical, since $Vol_k(t)$ cannot exceed the effective capacity and ideal metering tries to maintain $Vol_k^{met}(t)$ close to effective capacity. In fact, we have taken the effective capacity to be 3 percent below the maximum observed throughput. But see footnote 8 below.

For suppose only 30 vehicles could be accommodated at that ramp. This means that $120 - 30 = 90$ vehicles would have to be diverted away from that ramp, corresponding to less than 6 minutes of a 1,000 VPH ramp capacity. The diversion could be in time or in space or in mode.

The only calculation that remains is the one shown in Figure 2. That is now simple. The total number of VHT spent driving on the freeway section in period t is

$$\sum_k Vol_k^{met}(t) \times \frac{L_k}{60},$$

where L_k is the length of link MLk . This is the bottom graph of Figure 2. The middle graph is simply the sum of this and the waiting time at all the ramps.

5 Implications and caveats

The study presented above has implications for freeway performance measurement, planning, metering, and demand management. Some of the implications discussed below are more firmly supported than others.

Freeway performance

Travelers experience freeway performance by the the reliability of travel time. The latter could be estimated as a combination of mean and variance or the 70th or 90th percentile of the travel time distribution. To our knowledge, no agency calculates such distributions. PeMS does this routinely. Figure 28 shows the travel time distributions for a 78-mile trip beginning between 5 and 8 am on I-5N in Los Angeles for 20 weekdays in July 2000. An 80 percent confidence interval yields a travel time of between 60 and 105 minutes!

In the absence of travel time measurements, Caltrans uses congestion as a performance measure. It declares a link congested when speed drops below 35 mph for at least 15 minutes. The data presented here and a comprehensive study of data from 4,000 detectors in Los Angeles show conclusively that maximum flow occurs at 60 mph. So the only defensible measure for California is the one proposed here: congestion delay is the additional vehicle hours spent traveling below 60 mph. There is some support within Caltrans to adopt this measure.

Caltrans districts publish an annual congestion report, based on data from “floating” cars driven through 5-7 mile sections of freeways twice a year during congested periods [3]. Since the variation in travel times is enormous as Figure 28 indicates, these twice-a-year samples are unreliable. With a real-time system like PeMS it is now possible to track congestion accurately to determine trends as well as instantaneous departures from the trend. For example, PeMS can deliver a real-time alert whenever recurrent congestion exceeds the trend plus, say, twice the standard deviation.

Transportation agencies use Level of Service (LOS) as defined by the Highway Capacity Manual (HCM2000) [1, Chapter 23] as a freeway performance measure. HCM2000 gives a procedure to

calculate the speed and density on a freeway link, given the demand. A table defines LOS as a function of speed or density and the free flow speed. At the heart of the procedure are hypothesized speed-flow curves. In these curves the maximum throughput occurs at speeds well below free flow. For example, a link with a free flow speed of 62 mph supports a per lane flow of 1600 vehicles/hour at the free flow speed and a 50 percent larger maximum flow of 2400 vehicles/hour at 50 mph. Data from Los Angeles show that LOS and throughput calculated in this manner would be *completely wrong* since in every link maximum throughput occurs at 60 mph. For California at least, the HCM2000 speed-flow curves are invalid.

Capacity

The HCM formula defines the capacity of a link by its maximum throughput, which depends on the free flow speed through the above-mentioned speed-flow curves. The free flow speed itself may be either directly measured or determined using another formula that gives free flow speed as a function of the link's physical characteristics. Call the capacity calculated in this manner the **HCM capacity**. The HCM recommends using HCM capacity for operational uses (LOS calculation), and for design and planning (answering questions like the number of lanes needed to accommodate a certain flow at a specified LOS).

But as we saw there is little relation between a link's maximum observed throughput—its effective capacity—and the HCM capacity based on the link's physical capacity. Since both operational and planning decisions should be based on what a link can *actually* carry, it is effective capacity that should inform those decisions. Operational and planning decisions based on HCM capacity will usually be incorrect.

The study shows that effective capacity is an empirically stable notion. It remains an outstanding open question to explain this stability, and to calculate effective capacity on the basis of a link's connection to other links, the pattern of traffic, as well as its physical characteristics.

Lastly, effective capacity is *achievable* under the ideal metering policy defined here. It is an important open question whether there are metering policies that can achieve *sustained* throughput that exceed effective capacity.⁸

Ideal metering, excess demand, and system management

California freeway data suggest the idealized link behavior depicted in Figure 5. That behavior shows three regimes: until effective capacity is reached traffic is at free flow; beyond that lies congestion—the regime of decreasing speed and increasing density; finally, the recovery phase is reached when demand drops well below effective capacity.

The objective of ramp metering must be to achieve maximum flow and prevent the onset of congestion. The ideal ramp metering scheme is a local feedback rule that keeps occupancy downstream of

⁸This observation is due to Markos Papageorgiou. A close study of Figures 6-18 shows that in many links throughput increases by between 2 and 5 percent for a period of 5 minutes, just before congestion sets in. Can this increased throughput be sustained by appropriate ramp metering?

each ramp below a critical level. Such local feedback rules are well-understood [4]. The point made here is that the best critical level is where the flow reaches effective capacity.⁹

A practical ramp metering scheme must of course be more elaborate. The effective capacity will change over the course of the day (afternoon and morning commute hours) and week (weekends vs weekdays), and it will be affected by the weather. The critical level must correspondingly change. Second, the metering scheme must react to unanticipated changes, such as incidents or lane closures. Third, once a ramp metering scheme is in place, travelers will react to it, changing the pattern of demand and, as a consequence, the effective capacity might change. There should be a way to track these changes.

The ideal ramp metering scheme will lead to queues. When the queue at a ramp (Figures 19–26) becomes larger than can be accommodated there will be a spillover to city streets, provoking the ire of local citizens. Transportation officials deem such situations unacceptable, and ‘advanced’ metering schemes have over-rides that increase metering rates when queues become large. This, of course, is self-defeating as the resulting congestion will simply convert the nearby freeway into a parking lot, and the total delay will be much larger.

Suppose that the ramps in the study section can accommodate 30 vehicles. Then on-ramps 2, 4, 6, and 8 with maximum queue lengths of 170, 80, 120, and 40, respectively, will experience spillovers. Under ideal ramp metering the total number of vehicles that spill over during the entire congestion episode is the difference between the maximum queue length and the ramp capacity of 30, i.e. $140 + 50 + 90 + 10 = 290$ vehicles, amounting to less than two minutes of aggregate peak demand of 10,000 VPH. So instead of queue overrides, a more appropriate response is to divert traffic, through other means, including education. The procedures illustrated here can be used to calculate how much total delay is reduced at the cost of additional delay at the ramps. Moreover, changeable message signs that post ramp delays will encourage route divergence and relieve city streets.

The preceding two paragraphs underscore arguments for *corridor-wide* traffic management that coordinates the operation of ramp meters, arterial signals, changeable message signs, and other means of traffic control.

The calculations of the reductions in delay from metering and excess demand can be the starting point for the rational operation of a regional transportation system, because the calculations reveal the fundamental tradeoffs between freeway delay, queuing delay, and the distribution of demand over time and modes. Thus it is easy to answer questions like: how much diversion of demand originating at ramps in a given municipality is required to prevent spilling into city streets of queues at those ramps, or, how much additional delay will be caused if the traffic increased by, say, 5 percent. A quantitative evaluation of these tradeoffs can help assess the effectiveness of different transportation options: additional ramp storage capacity, an extra lane on a link, transit improvement, providing traveler information, etc.

Many California freeways have a HoV (High-occupancy Vehicle) lane on which one can legally drive a vehicle during congestion periods provided it carries at least two or three persons. HoV lanes enjoy free flow conditions during congestion. But usually the flow (in VPH, not in persons per

⁹The ideal local feedback rule will perform better than the sophisticated system-wide adaptive controls of the SWARM system being installed in California TMCs.

hour) is below maximum throughput. If perfect metering is implemented, all lanes would experience free flow, and the use of HoV lanes would become problematic.

Measuring non-recurrent congestion

This important topic is only briefly discussed here. A more extended discussion will be presented in the future, following serious empirical study.

Non-recurrent congestion is supposed to account for a significant proportion of delays and additional accidents. But these claims lack empirical basis. The first difficulty is that recurrent congestion is a random quantity, varying from one day to the next (see Table 1); non-recurrent congestion is even more random. Furthermore, recurrent and non-recurrent congestion sometimes occur together. Thus it is very difficult to separate out the contributions of the two causes of congestion.¹⁰ The second difficulty is the absence of reasonably complete datasets that include both traffic measurements and incidents. The PeMS system, when it is augmented with incident and lane-closure data, will take care of the second difficulty.¹¹

The congestion measure proposed here can address the first difficulty. Consider again the study section and the morning congestion. We know the statistical distribution of the recurrent congestion. If there is an incident in this section during this period, PeMS can calculate the total congestion (the result of both recurrent and non-recurrent congestion). If congestion data for many incidents are available, we would create a statistical distribution of this total congestion. Comparing it with the distribution of the incident-free recurrent congestion will reliably estimate the contribution of incidents.

6 Conclusions

Freeway operational and planning decisions today rely on theories and practices exemplified, for example, by the Highway Capacity Manual, and simulation packages like Corsim or Paramics. PeMS studies show that these theories and practices are based on crucial hypothesized relationships—between speed and flow, and between a link’s capacity and its physical characteristics—that are in fact wrong. With access to large data sets provided by PeMS, the work needed to calibrate simulation models can be automated to a considerable extent. Simulation will then become a reliable tool for both real-time operations and long-term planning studies.

PeMS is a large-scale freeway data collection, storage, and analysis project. It provides real-time and historical information of use to managers, engineers, planners, researchers, and travelers. It is inexpensive, easy to maintain, and can be readily duplicated in other states. PeMS makes available information based on masses of empirical data that can be used to create a reliable freeway traffic theory, a trustworthy practice of traffic engineering, and well-informed public transportation policy choices.

¹⁰See the careful I-880 Freeway Service Patrol study [5].

¹¹PeMS now collects incident data published by the California Highway Patrol. These data are not yet integrated with the loop-detector data.

References

- [1] Transportation Research Board. *Highway Capacity Manual 2000*. National Research Council, 1998.
- [2] C. Chen, Z. Jia, K. Petty, A. Skabardonis, and P. Varaiya. Freeway performance measurement system: mining loop detector data. To be presented at TRB Annual Meeting, January 2001. Available at www.path.berkeley.edu/~varaiya/papers_ps.dir/pems_paperf.pdf.
- [3] Office of Highway Operations. Highway congestion monitoring report. Technical report, Caltrans-District 4, 1998.
- [4] M. Papageorgiou, H. Hadj-Salem, and J. Blosseville. Alinea: a local feedback control law for on-ramp metering. *Transportation Research Record*, 1320, 1991.
- [5] A. Skabardonis, K. Petty, H. Noeimi, D. Rudzewski, and P. Varaiya. I-880 field experiment: Database and incident delay procedures. *Transportation Research Record*, 1554, 1996.

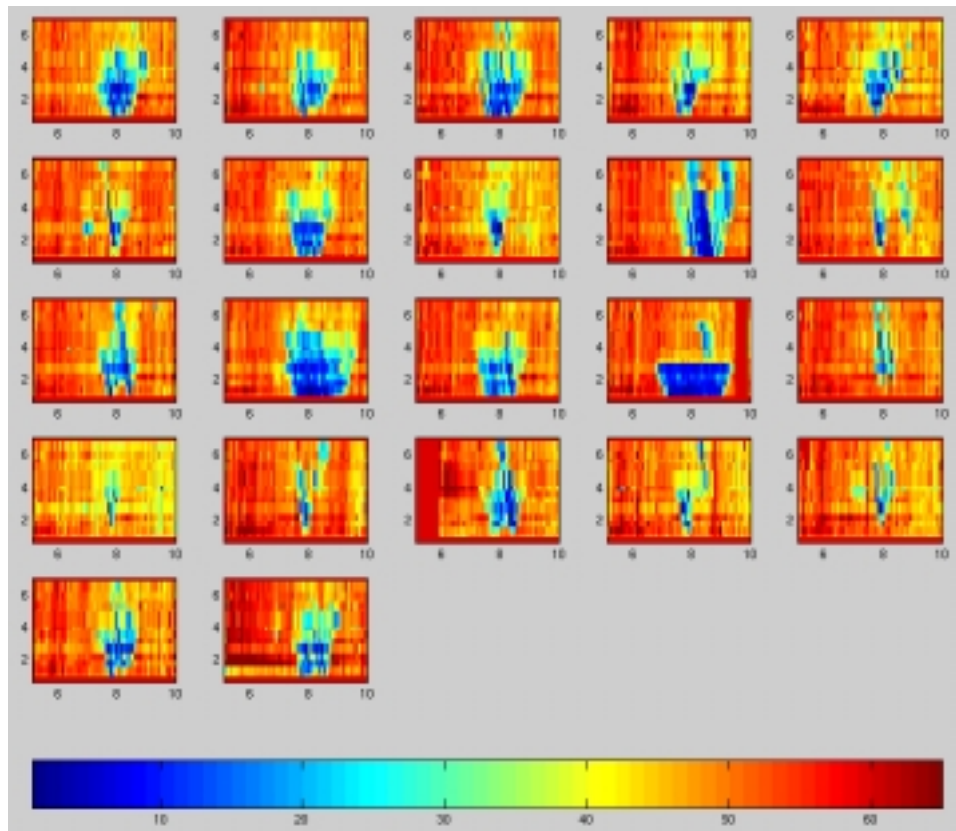


Figure 1: Contour plot of speed for a section of I-405N, for 22 weekdays in June 1998. The x-axis is time, from 5.00 to 10.00 am; the y-axis is distance from postmile 0 to 8. Vehicles travel from bottom to top. Darker points correspond to lower speed. The chokepoint link is near postmile 5.

7 Figures

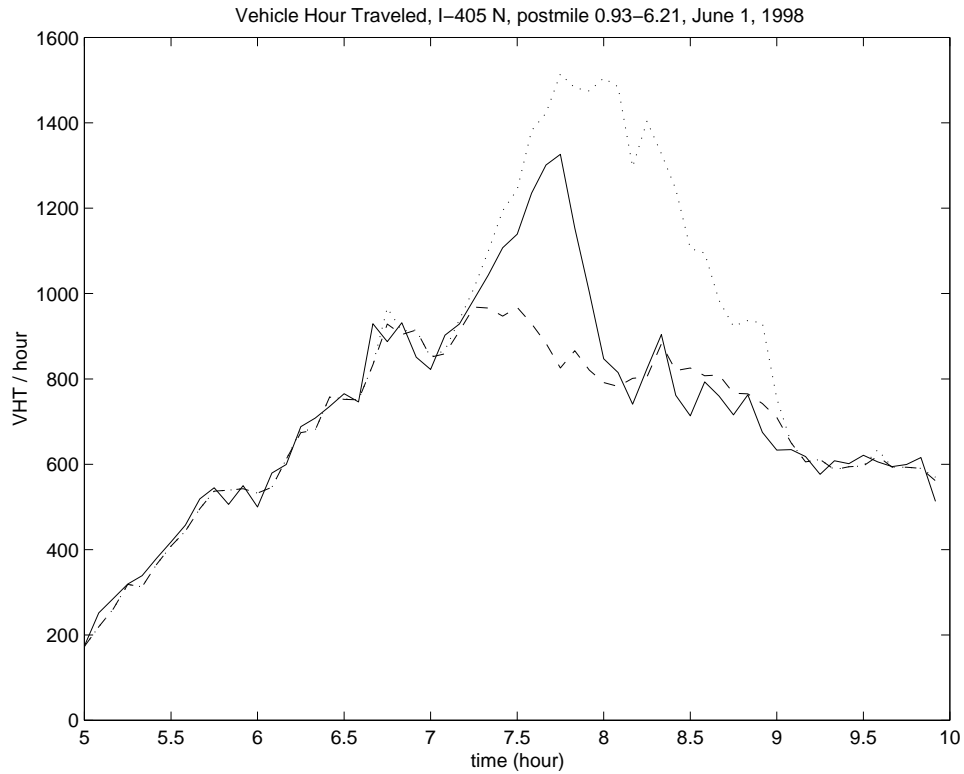


Figure 2: The top graph is the amount of time in VHT actually spent on the freeway section, every 5 minutes. The units are normalized to VHT per hour, so the the total vehicle-hours spent on this section, between 5.00 and 10.00 am is the area under the top graph. The middle graph is the VHT per hour under ideal metering, including time on the ramps. The bottom graph excludes time spent on the ramps, so it is the VHT per hour that would be spent traveling at 60 mph.

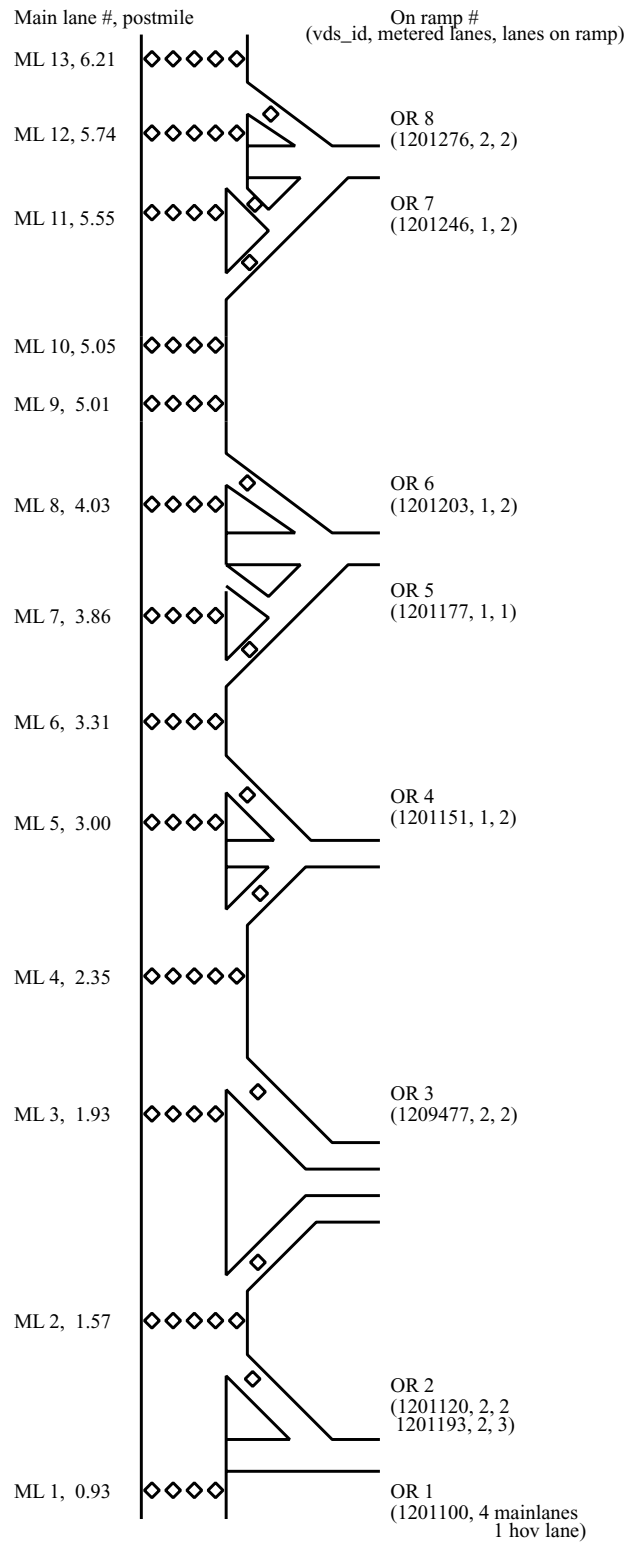


Figure 3: The freeway study section has eight on and off-ramps.

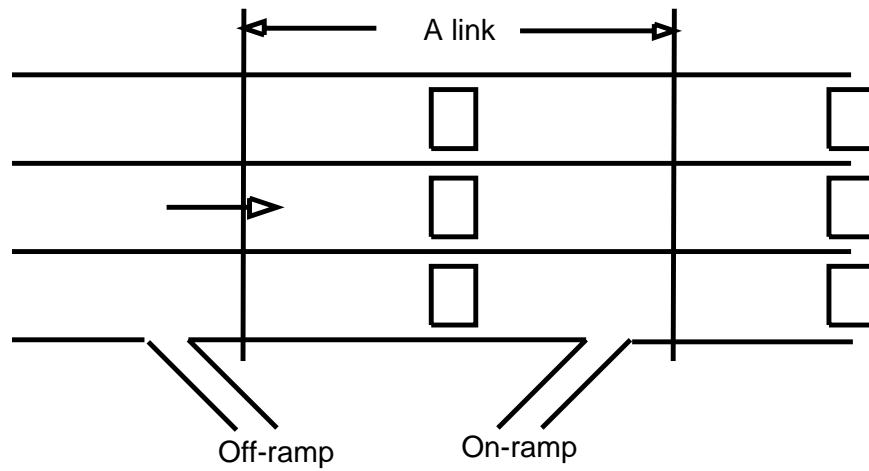


Figure 4: A link is a portion of freeway associated with a set of loops, and may contain one on- or off-ramp.

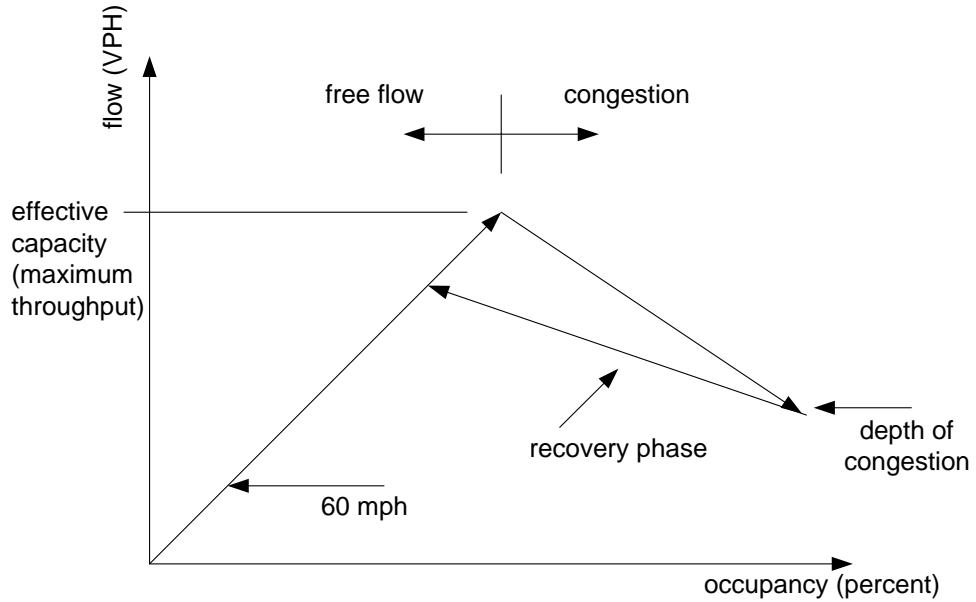


Figure 5: Typical link behavior during a congestion episode. Vehicles travel at 60 mph until flow reaches capacity. Congestion starts. Speed and flow drop, occupancy increases to a maximum value. Demand then drops and speed gradually increases.

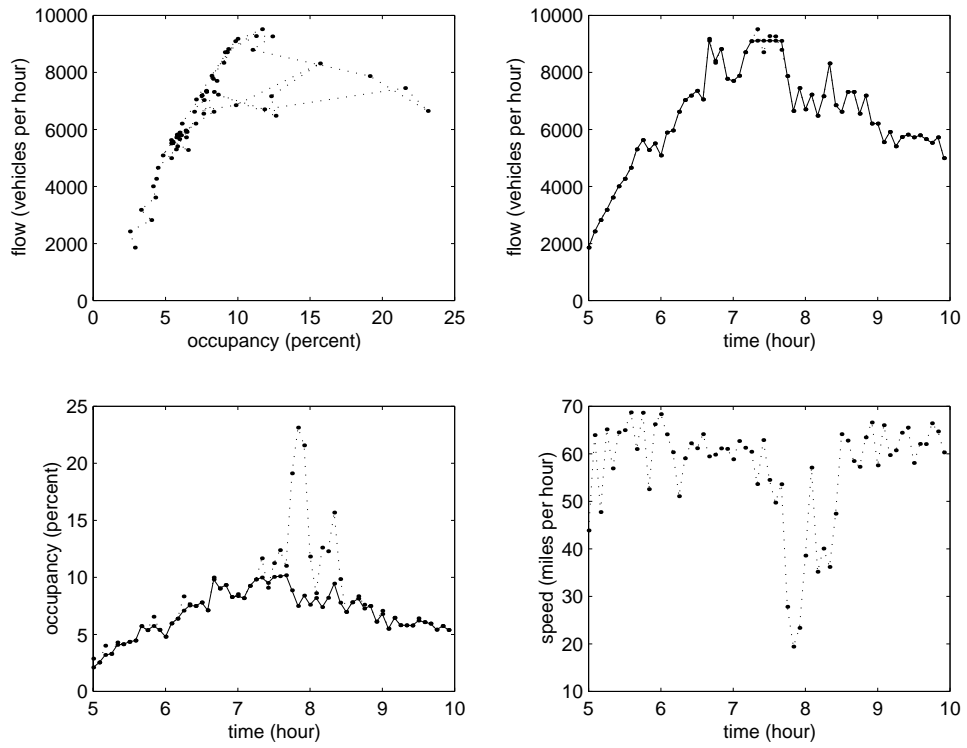


Figure 6: Performance of links ML1: actual and ideal metering.

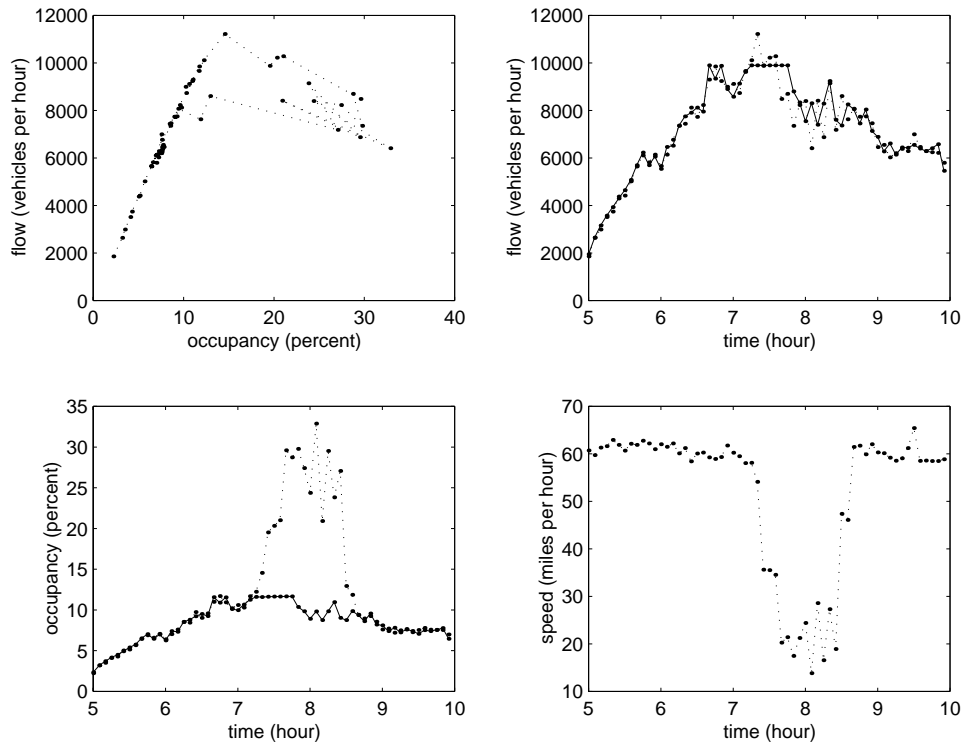


Figure 7: Performance of links ML2: actual and ideal metering.

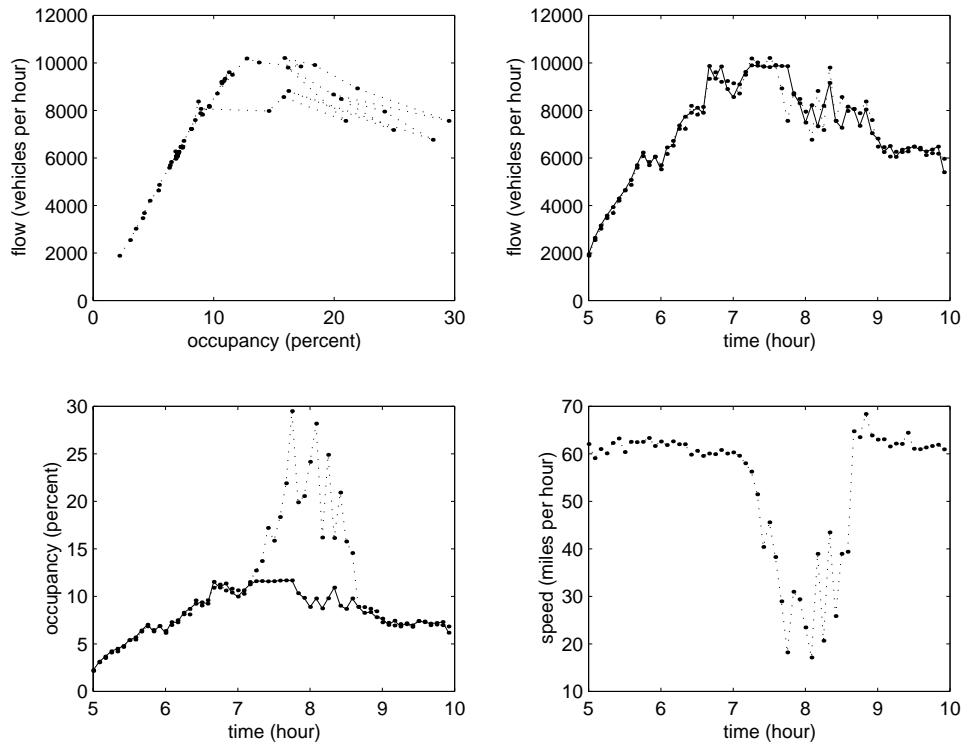


Figure 8: Performance of links ML3: actual and ideal metering.

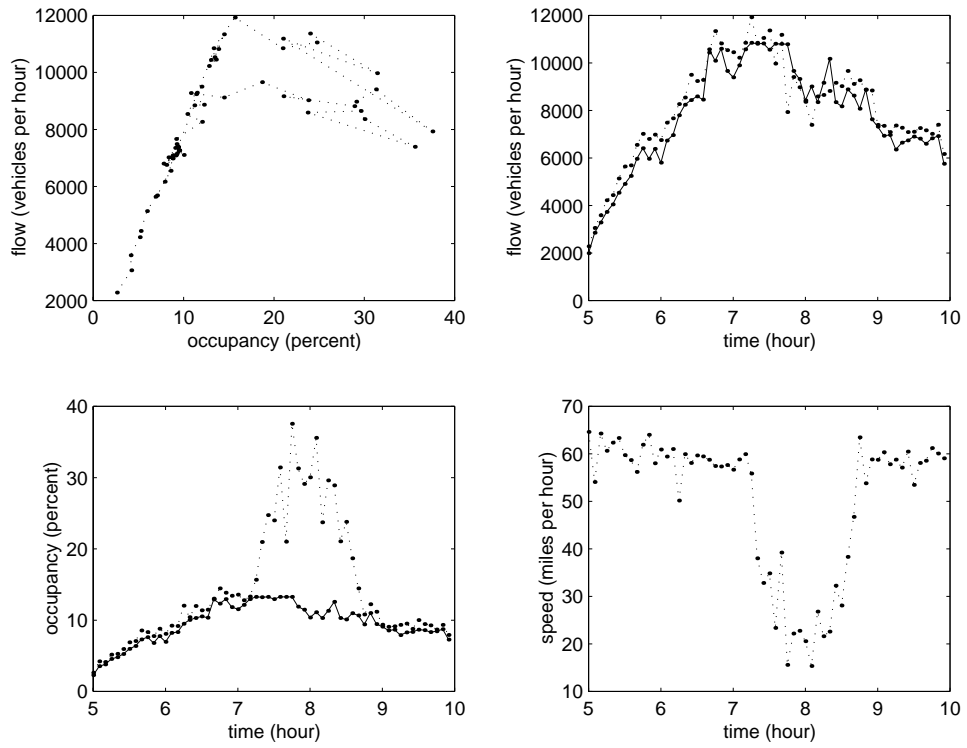


Figure 9: Performance of links ML4: actual and ideal metering.

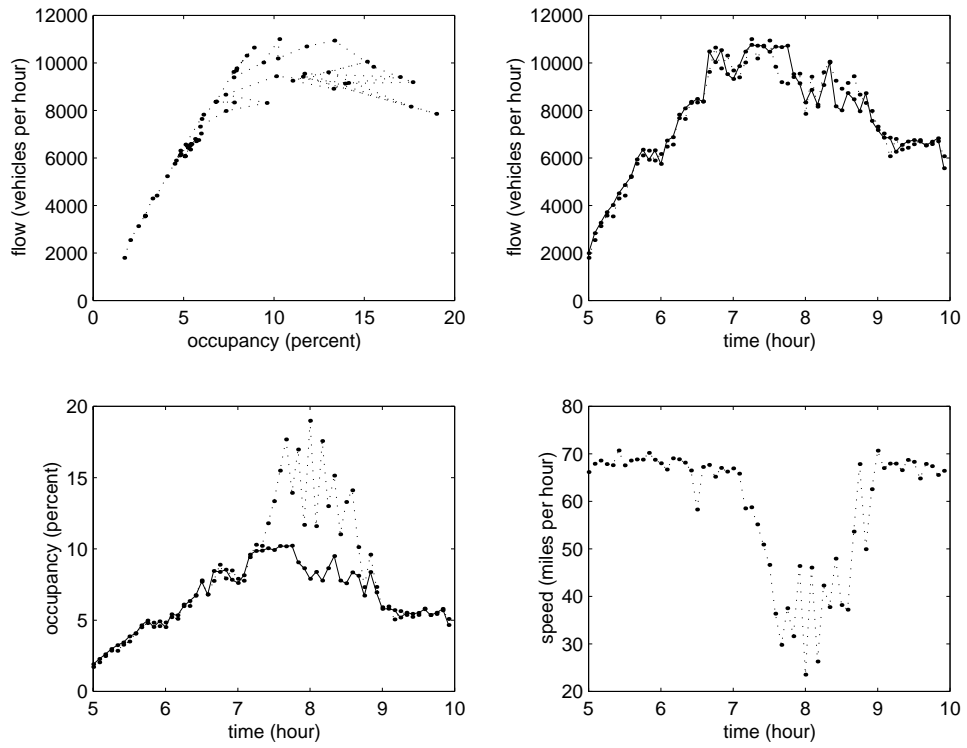


Figure 10: Performance of links ML5: actual and ideal metering.

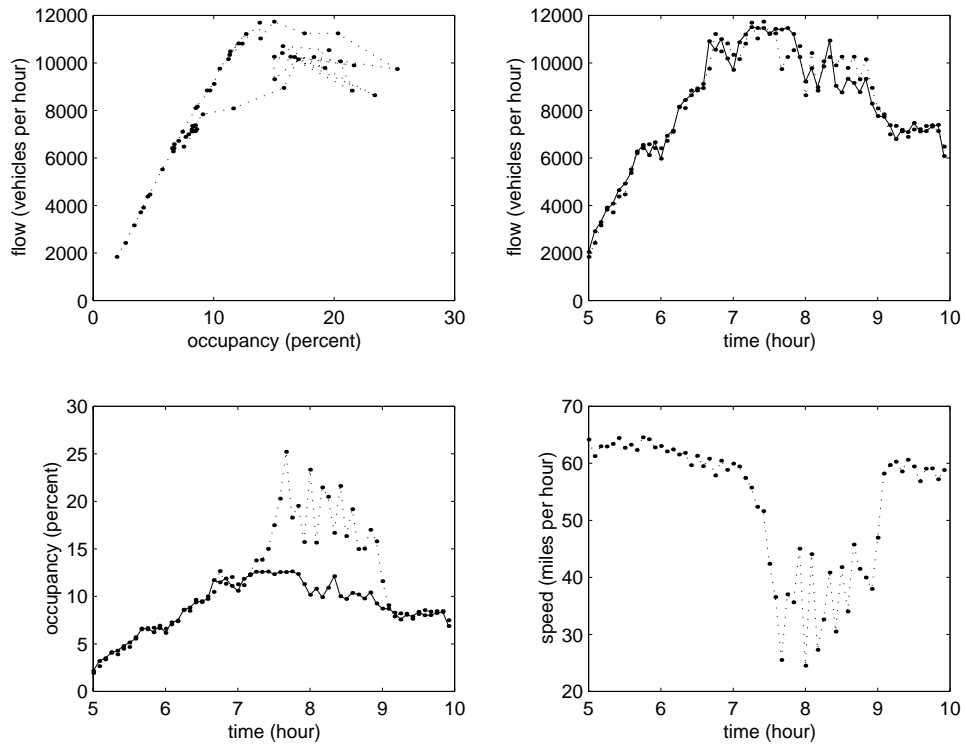


Figure 11: Performance of links ML6: actual and ideal metering.

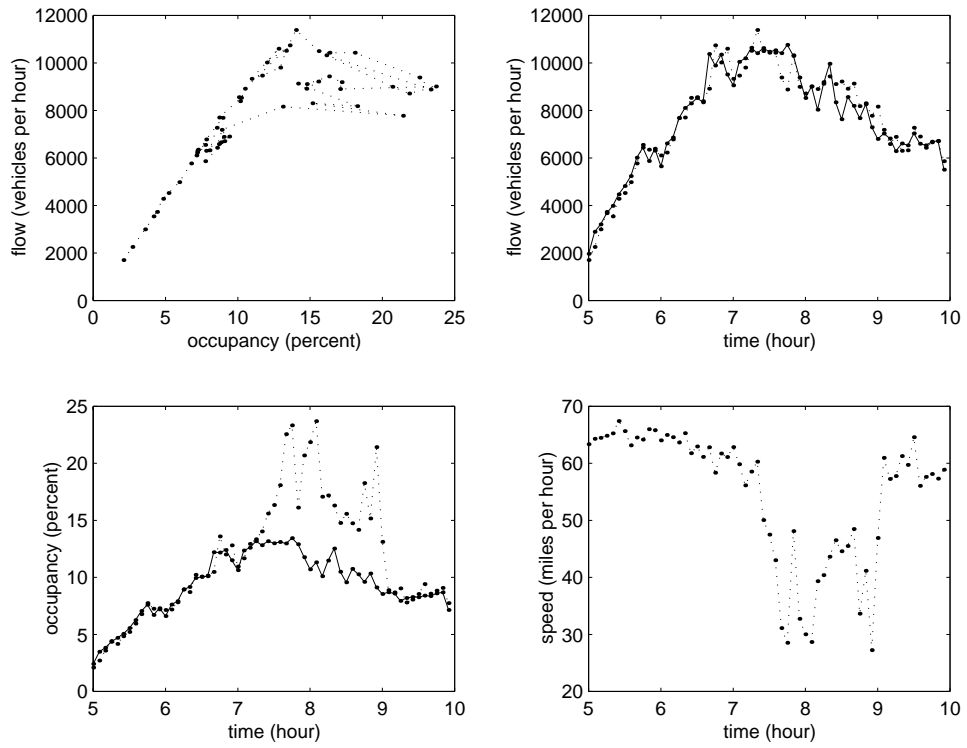


Figure 12: Performance of links ML7: actual and ideal metering.

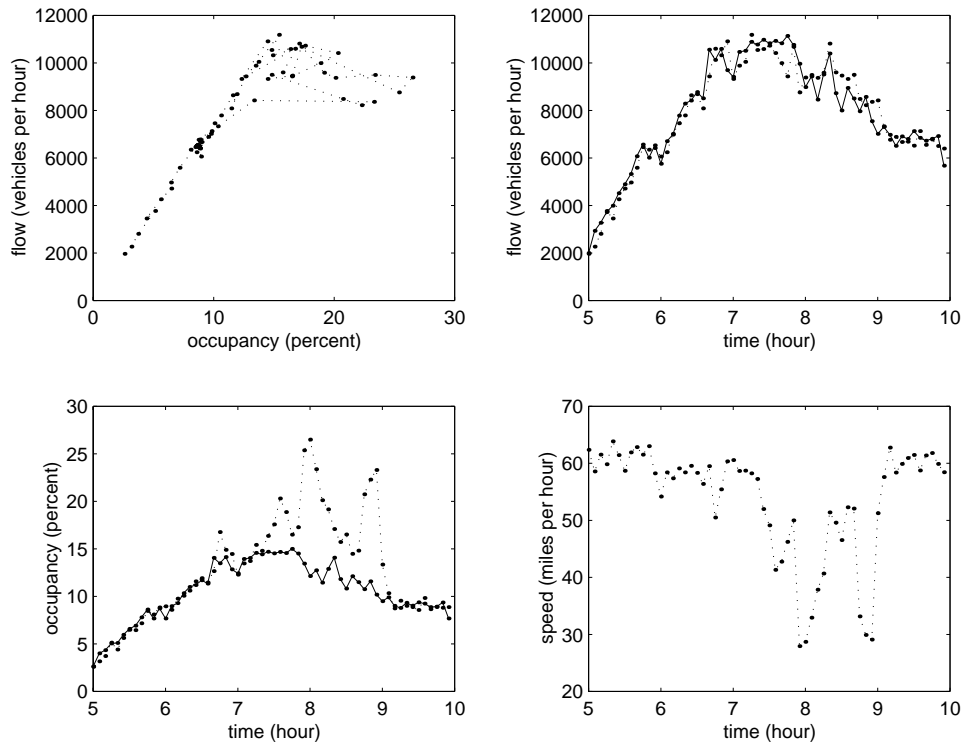


Figure 13: Performance of links ML8: actual and ideal metering.

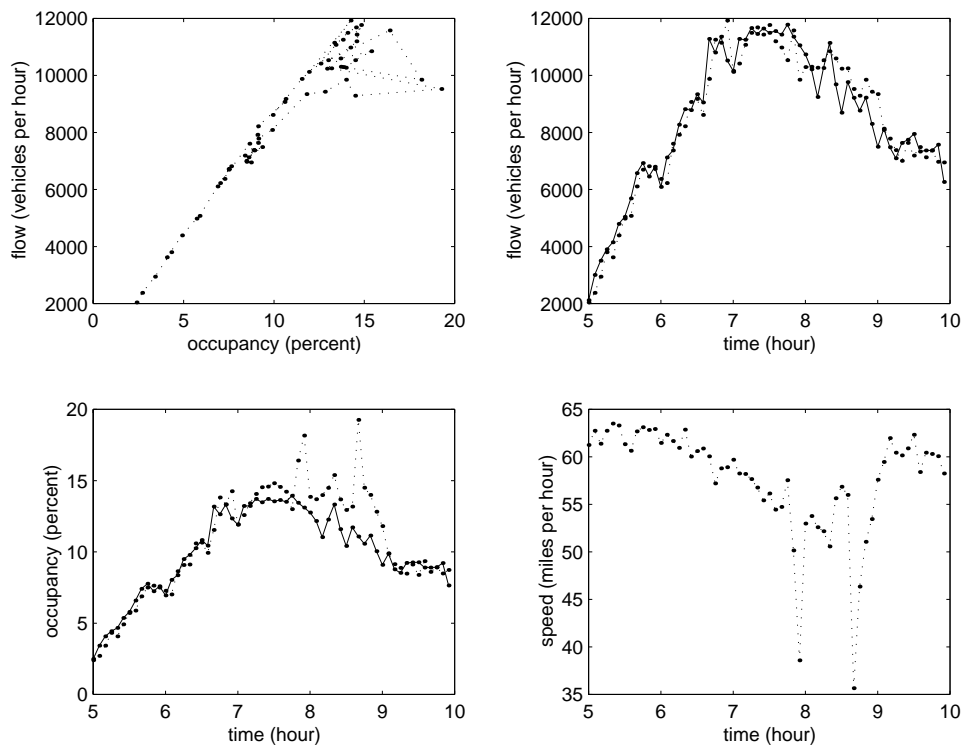


Figure 14: Performance of links ML9: actual and ideal metering.

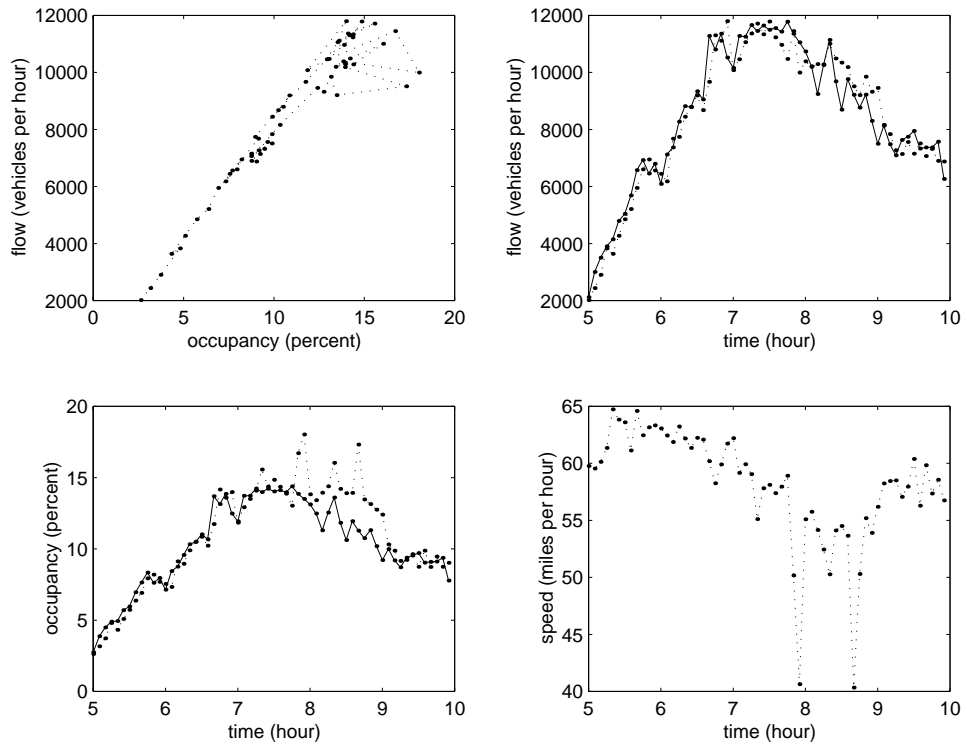


Figure 15: Performance of links ML10: actual and ideal metering.

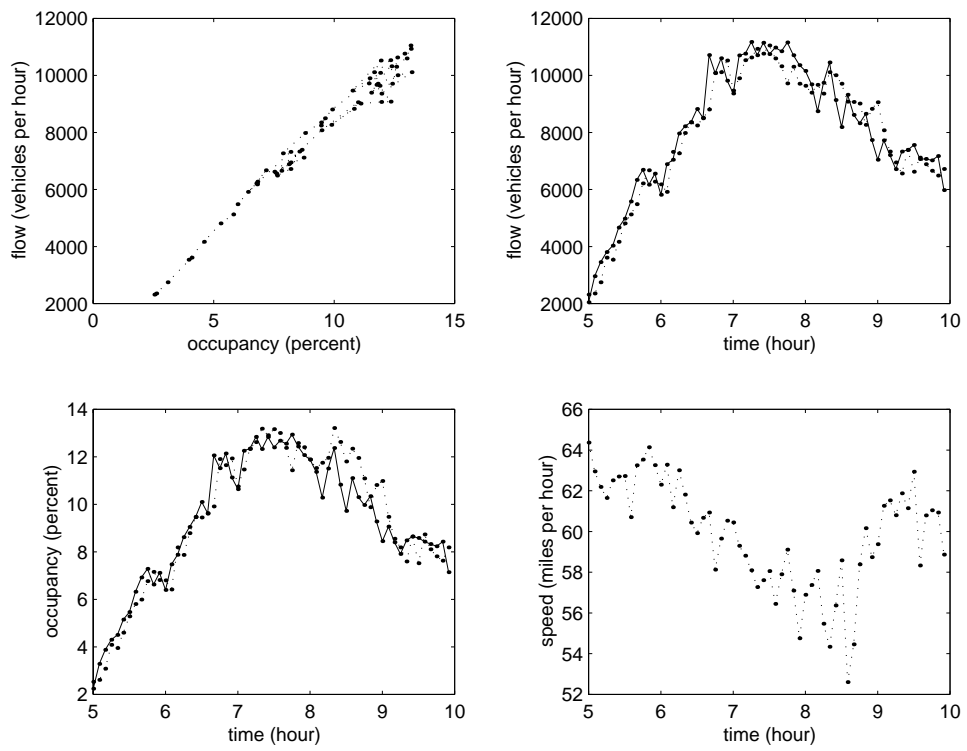


Figure 16: Performance of links ML11: actual and ideal metering.

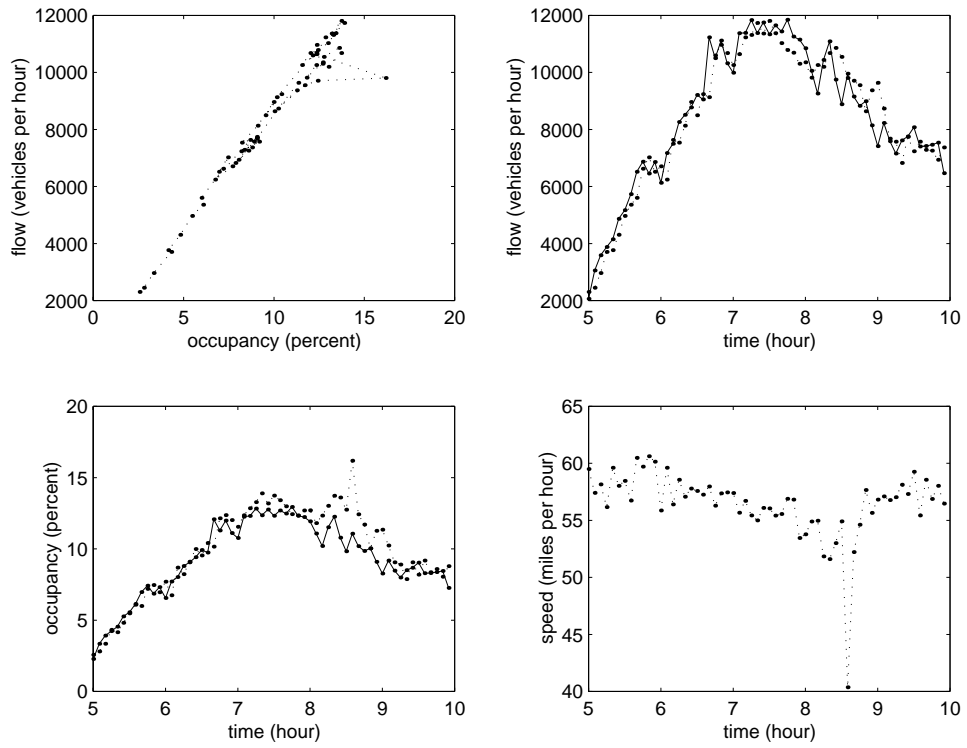


Figure 17: Performance of links ML12: actual and ideal metering.

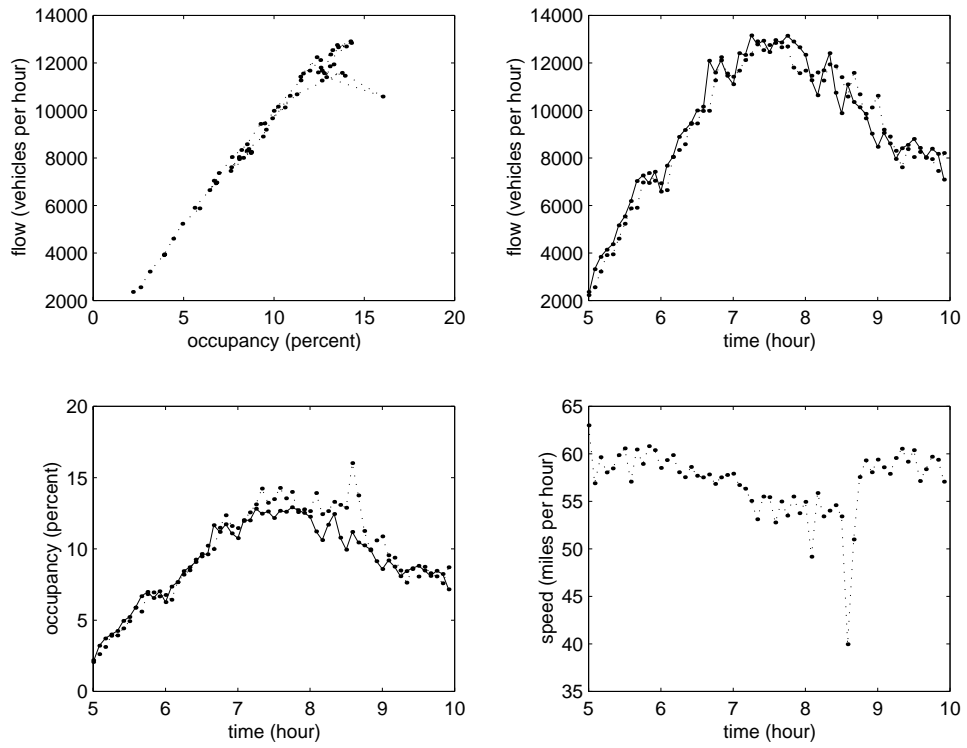


Figure 18: Performance of links ML13: actual and ideal metering.

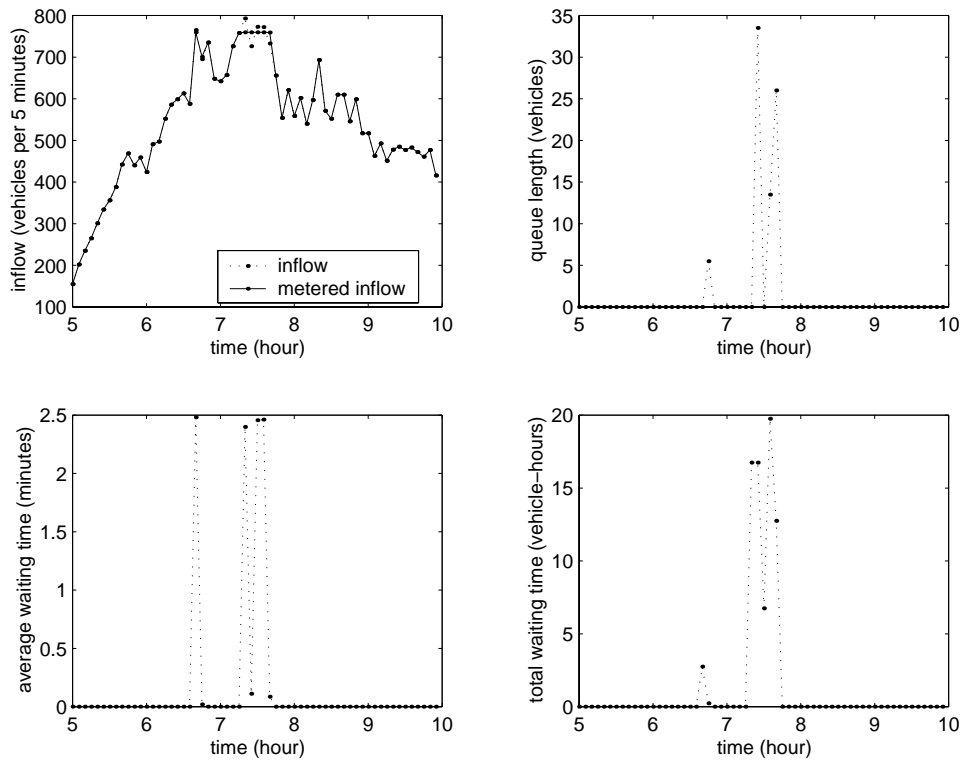


Figure 19: Queue behavior at (virtual) on-ramp 1 under metering. The inflow here is the upstream flow into the study section.

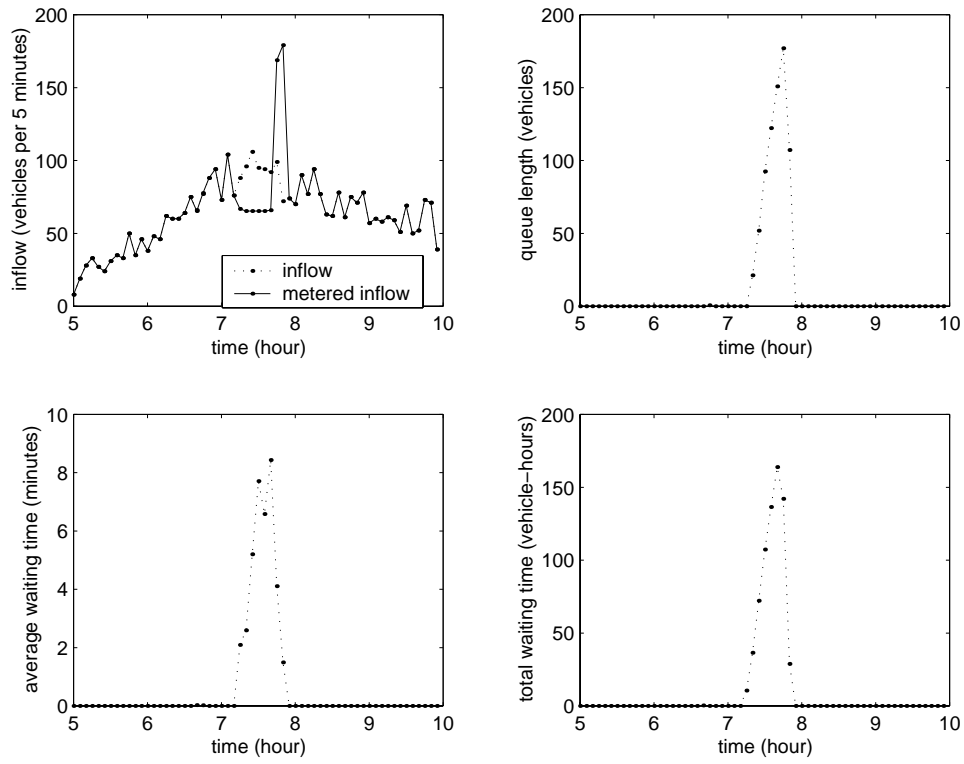


Figure 20: Queue behavior at on-ramp 2 under metering.

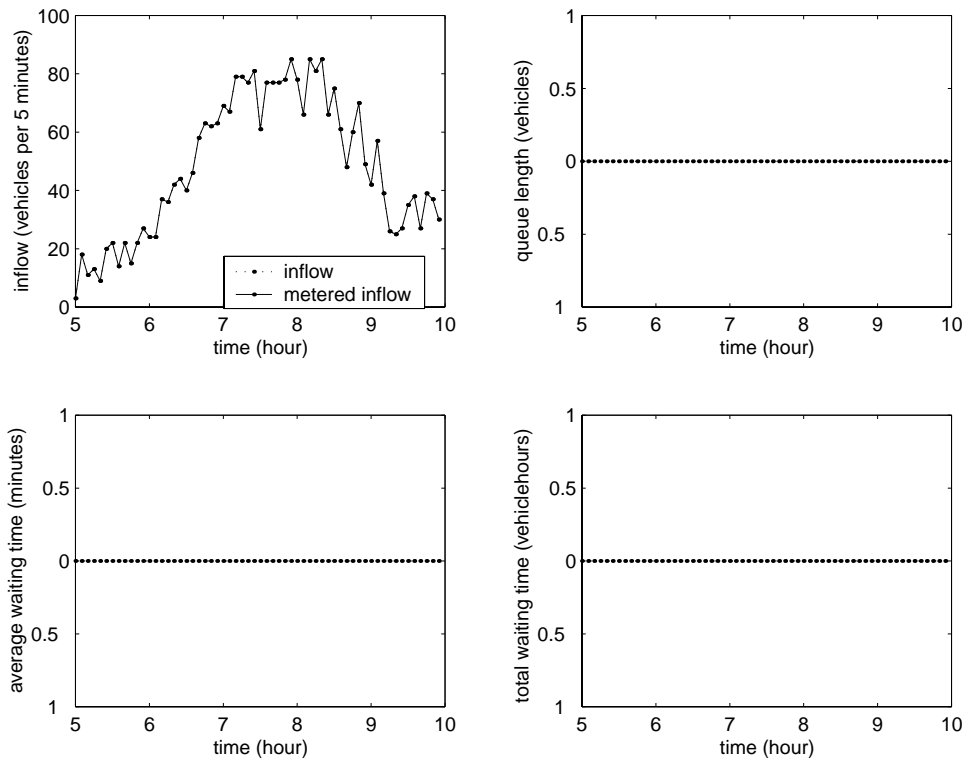


Figure 21: Queue behavior at on-ramp 3 under metering.

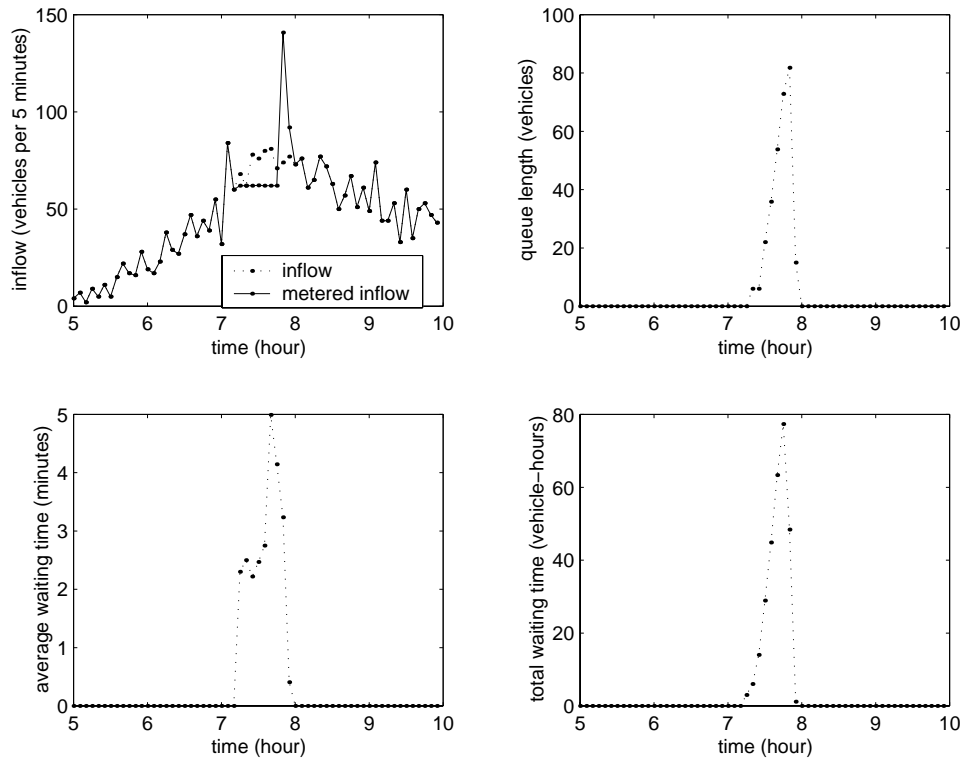


Figure 22: Queue behavior at on-ramp 4 under metering.

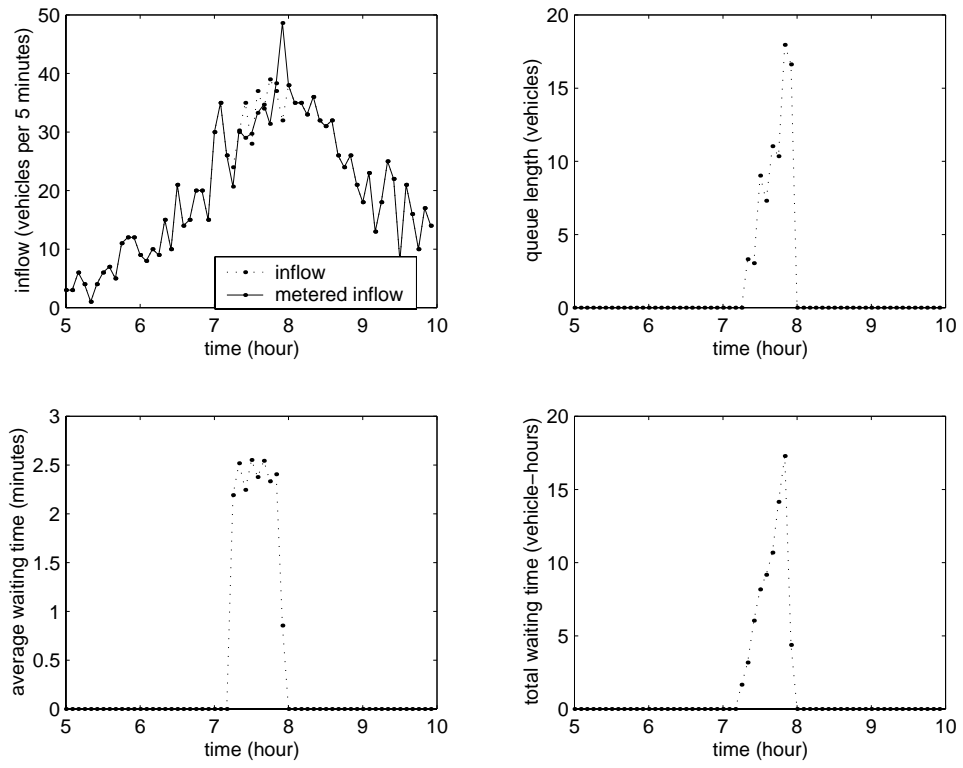


Figure 23: Queue behavior at on-ramp 5 under metering.

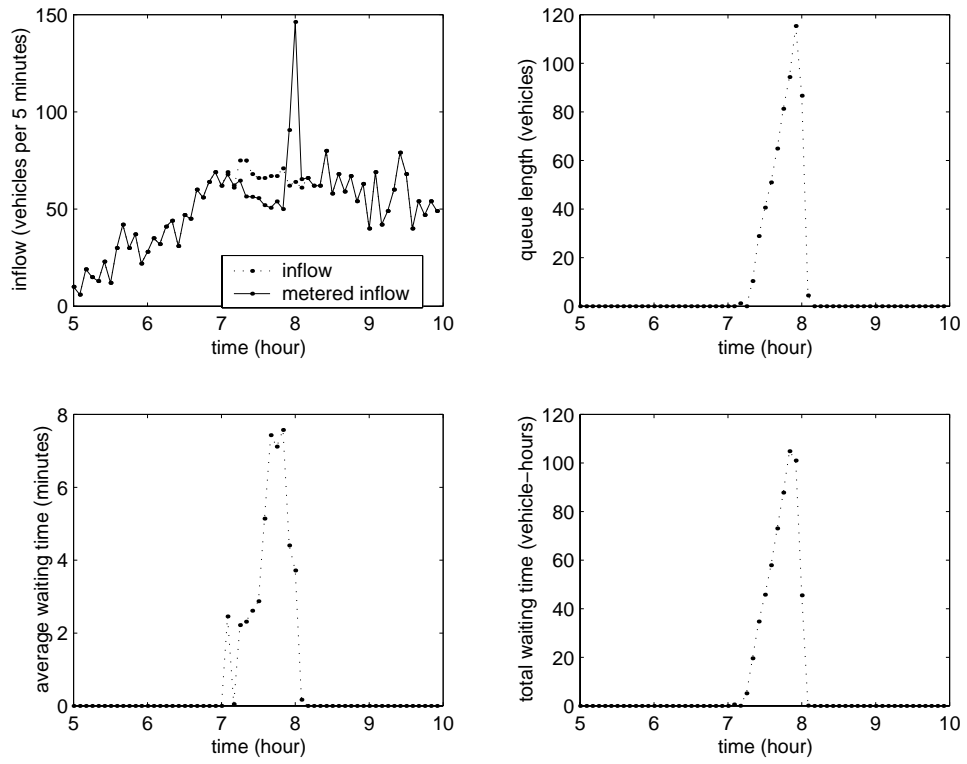


Figure 24: Queue behavior at on-ramp 6 under metering.

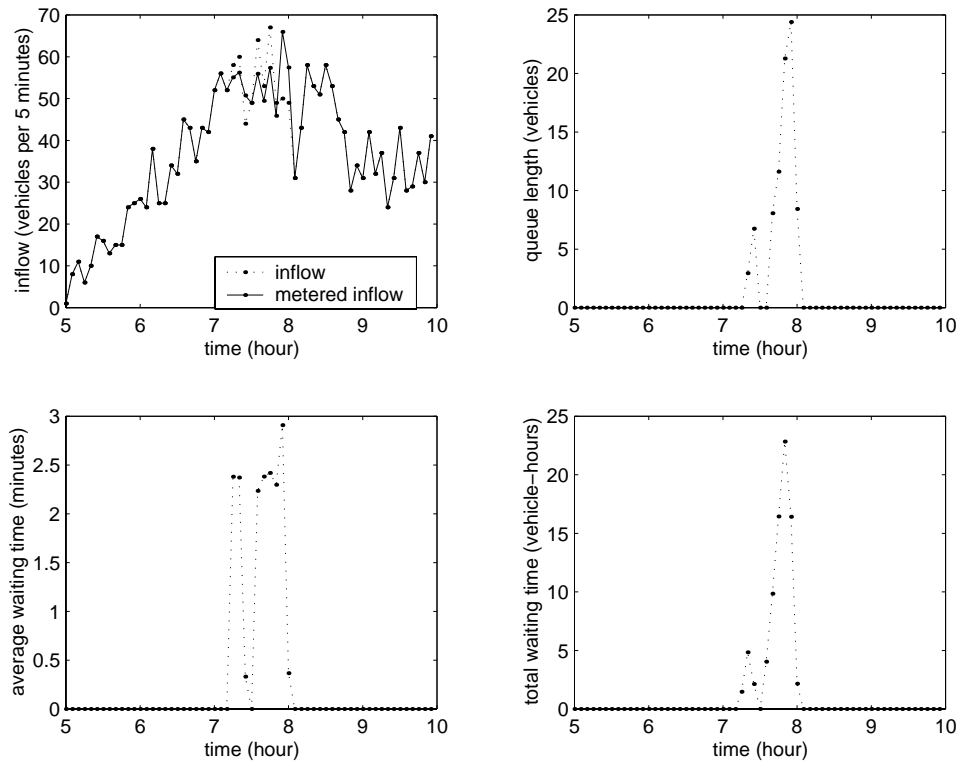


Figure 25: Queue behavior at on-ramp 7 under metering.

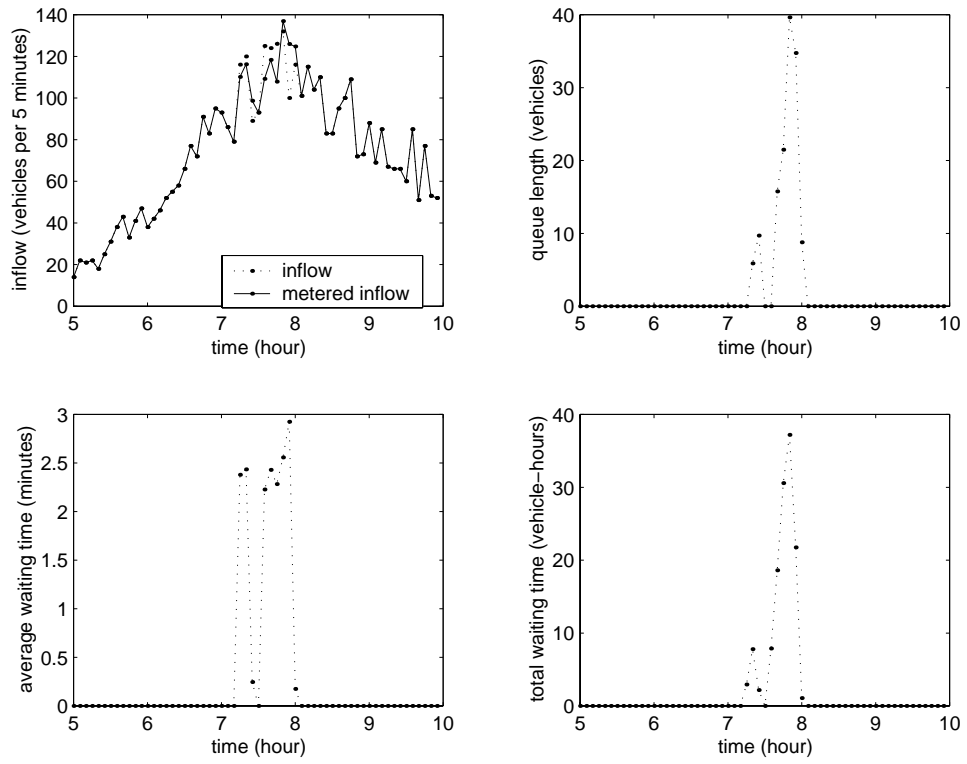


Figure 26: Queue behavior at on-ramp 8 under metering.

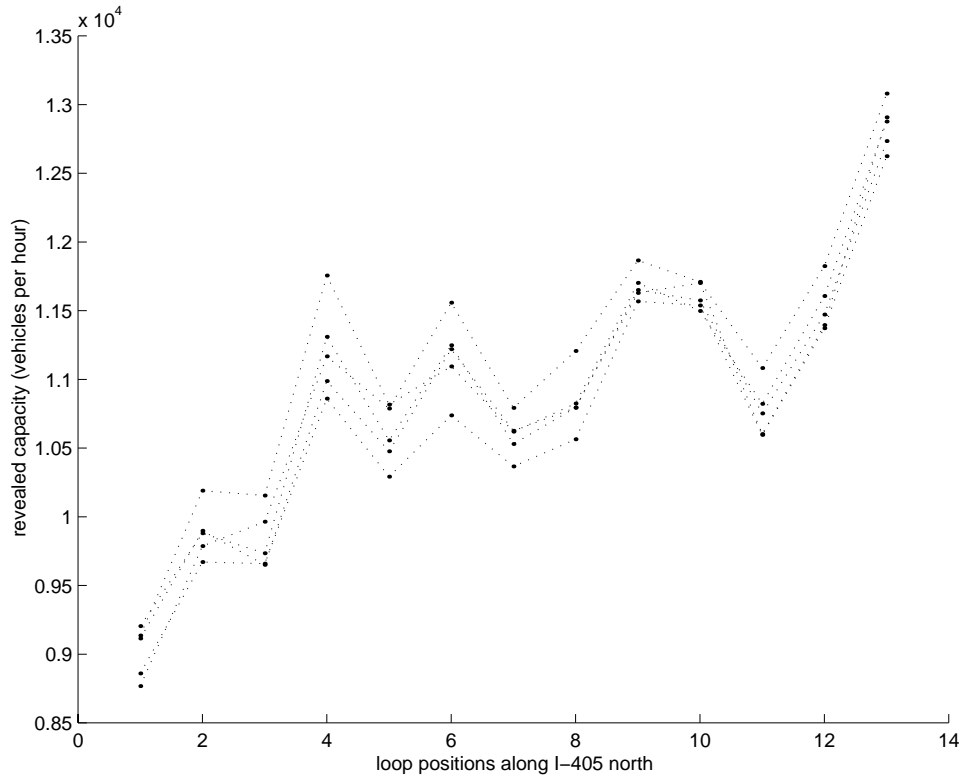


Figure 27: The PeMS capacity calculations for the links in the study section for five days in June 1998. The close agreement lends credence to the PeMS capacity definition.

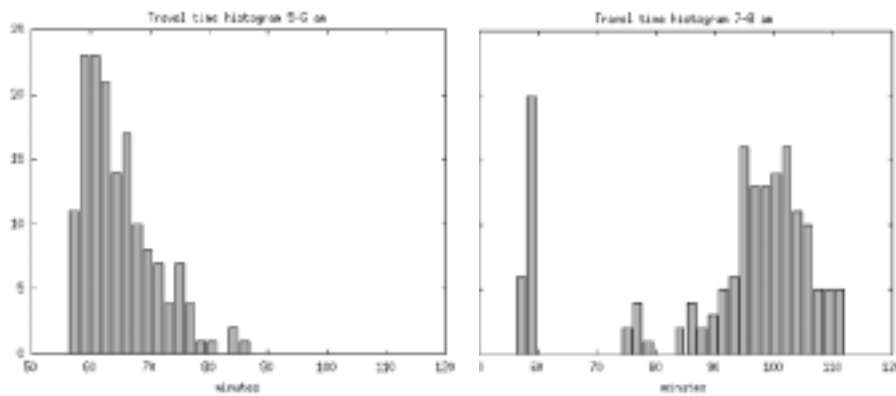


Figure 28: Travel time distributions for a 78-mile trip on I-5N in Los Angeles during weekdays in July 2000. The figure on the left (right) is for trips starting between 5.00 and 6.00 am (7.00 and 8.00 am). The outliers on the right are for July 3,4 holidays.