

1
2
3
4
5
6
7
8
9
10
11
12 **THE CONGESTION PIE: DELAY FROM COLLISIONS, POTENTIAL RAMP**
13 **METERING GAIN, AND EXCESS DEMAND**
14

15
16 **Jaimyoung Kwon¹**

17 Institute of Transportation Studies
18 University of California, Berkeley
19 367 Evans Hall, Berkeley, CA 94720-3860
20 Tel: (510)642-2781; Fax: (510)642-7892
21 kwon@stat.berkeley.edu
22

23 **Pravin Varaiya**

24 Department of Electrical Engineering and Computer Science
25 University of California, Berkeley, 94720-1770
26 Tel:(510)642-5270; Fax:(510)643-7815
27 varaiya@eecs.berkeley.edu
28
29
30

31 Transportation Research Board
32 84th Annual Meeting
33 January 2005
34 Washington, DC
35

36
37 4200 words (excluding Figure captions)
38 Plus 7 Figures (1750)
39 total 5950 words
40
41
42
43
44
45
46
47

48

¹Corresponding author
49
50
51
52

ABSTRACT

A method is presented to divide the total congestion in any freeway section into three 'slices': (1) the delay caused by collisions; (2) the potential reduction in delay that can be achieved by ramp metering at bottlenecks; and (3) the remaining delay, most of which is attributed to demand that exceeds the maximum sustainable flow. The method involves two steps. The first step determines the space-time region affected by a collision and then estimates the additional delay caused by the collision. The second step identifies bottlenecks and then estimates the potential reduction in delay that could be achieved by appropriate ramp metering. The remaining delay is the residual. This is largely due to excess demand, but also includes all other causes, including non-collision incidents, lane closures, and weather. The method is fully automated and can be applied to any site with minimum calibration. It requires traffic volume and speed data, and the time and location of each collision occurrence. Applied to a 22.5-mile section of I-15 in San Diego, the method reveals that collisions, potential reduction by ramp metering, and excess demand respectively account for 31%, 46%, and 23% of the total daily delay. The large value (46%) signals the great potential to mitigate congestion by ramp metering. In addition to the three congestion pie slices, the method estimates the probability distributions of duration, spatial extent, and additional delay caused by individual collisions, which reveal, for example, that the delay impact of a collision depends significantly on traffic conditions.

1 INTRODUCTION

Congestion has many causes, including inefficient operations, collisions, excess demand, lane closures, and weather. Their impact can be summarized in the division of the congestion ‘pie’ into its constituent slices as in Figure 6. Knowledge of the congestion pie can be used to select effective congestion mitigation strategies at a particular site, and to allocate resources to implement those strategies.

The paper presents a method to divide the total congestion D_{tot} in any freeway section into three slices: (1) D_{col} , the congestion caused by collisions, which could be reduced by, say, a quicker incident response system; (2) D_{pot} , the congestion that potentially can be eliminated by appropriate ramp metering at bottlenecks; and (3) the residual delay, D_{rem} , that can largely be attributed to demand that exceeds the maximum sustainable flow, but also includes congestion caused by other factors such as lane closures and weather. Thus the method refines previous studies which group D_{pot} and D_{rem} together as ‘recurrent’ congestion.

Transportation agencies typically measure and report recurrent congestion, which accounts for 40%-70% of total congestion. The availability of more comprehensive data has prompted attempts to estimate the relative contributions of different causes of congestion. There are studies that divide total congestion into ‘recurrent’ and ‘non-recurrent’ congestion; and studies that divide the non-recurrent congestion into accident-induced congestion and other incident-induced congestion. There also are estimates of the congestion caused by adverse weather. These studies are reviewed in the next section.

All these studies leave unexplained a large fraction (between 40 and 70 percent) of the total congestion. This residual is often called ‘recurrent’. As Hallenbach et al. observe “Many large delays still occur for which incidents are not responsible, and for which no ‘cause’ is present in the [data].” They suggest that one cause of these delays may be “unusual volume surges at ramps ... that are not being effectively handled by the ramp metering program” (1, p.11). The proposed method estimates this potential reduction in delay, D_{pot} .

Application of the method to a 22.5-mile section of northbound I-15 in San Diego indicates that 46% of the total congestion delay could be eliminated by appropriate ramp metering. Improved operations can potentially achieve this very large delay reduction. From a management perspective, this unrealized gain is equivalent to a loss in productivity. The System Metrics Group analysis of the performance of Los Angeles freeways finds large productivity losses (2, pp. 23-29). The analysis defines the percent productivity of a congested segment as the actual flow divided by 2000 (the nominal per lane capacity). Productivity losses in the most congested segments are combined by weighting each segment with its length and congestion duration, to yield a productivity loss equivalent to 200 lane-miles during the PM peak.

The rest of the paper is organized as follows. Previous studies are reviewed in Section 2. The proposed method is described in Section 3. The congestion pie for the I-15, San Diego site is constructed in Section 4. Section 5 concludes the paper.

2 PREVIOUS STUDIES

Transportation agencies until recently only reported recurrent congestion (e.g. (3)). The availability of more comprehensive data has prompted efforts to quantify the relative impact of different causes of congestion.

Several studies estimate the impact of incidents. The earliest studies relied on correlating specialized collection of incident data in the field using ‘floating cars’ with loop-detector based traffic data, in order to conduct a cost-benefit study of Freeway Service Patrol (FSP), (4). These incident data collection efforts provide a great deal of information about the nature of incidents (e.g. fewer than 4 percent of FSP assists deal

with accidents), but they are too expensive to replicate on a large scale or on a continuing basis. Moreover, significant manual intervention is needed to analyze the field data.

California Highway Patrol computer aided dispatch (CAD) logs and the FSP logs provide large data sets. These have been used for Los Angeles freeways (5), and Oregon (6). However, the CAD data have to be manually processed. The studies cited above are concerned with estimating the performance of incident response measures like the FSP.

An important element of the manual processing is the determination of the spatial and temporal extent of the congestion impact of an incident. A significant contribution of the method proposed in this paper is the algorithm that automates the delineation of a collision's impact.

Determining an incident's impact region can be avoided if one is willing to average out the impact of individual incidents as is done in the studies by Skabardonis et al (7) and Hallenbach et al (1). The two studies separate 'non-recurrent' and 'recurrent' congestion; they differ both in terms of definition and method.

The study in (7) considers a freeway section for a peak period. The total congestion on each of several days is defined as the additional aggregate vehicle-hours traveled, driving below 60 mph (see (1) below). Each day is classified as 'incident-free' or 'incident-present'. The average total congestion in 'incident-free' days is defined to be the recurrent delay. Total congestion in 'incident-present' days is considered to be the sum of recurrent and incident-induced congestion. Subtracting average recurrent congestion from this gives an estimate of the average non-recurrent or incident-induced congestion. Similarly, (1) defined the median traffic conditions on days when a freeway section does *not* experience lane-blocking incidents as the "expected, recurring condition."

A less data-intensive approach is suggested by Bremmer et al (8). In the absence of incident data, they infer that an incident has occurred if a trip "takes twice as long as a free-flow trip for that route." The purpose of the study is to forecast travel times, measure travel time reliability, and cost-benefit analysis of operational improvements, rather than to measure congestion impact of different causes.

The effect of inclement weather on freeway congestion is studied in (9, Chapter 22) and (10). According to those studies, light rain or snow can reduce traffic speed by roughly 10 percent, heavy rain can decrease highway speeds by approximately 16 percent and in heavy snow, freeway speeds can decline by about 40 percent. Lastly, the FHWA website (11) combines various studies to produce a rough breakdown of the impact of several congestion causes.

3 PROPOSED METHOD

The study site is a contiguous section of freeway with n detectors indexed $i = 1, \dots, n$, each providing flow (volume) and speed measurements averaged over 5-minute intervals indexed $t = 1, \dots, T$. Detector i is located at postmile x_i ; $v_i(t) = v(x_i, t)$ is its measured speed (miles per hour, mph) and $q_i(t) = q(x_i, t)$ is its measured flow (vehicles per hour, vph) at time t . If $x_i < x_j$, it is understood that x_i is upstream of x_j .

The n detectors divide the freeway section into n segments. Each segment's (congestion) delay is defined as the difference between vehicle-hours actually traveled and the minimum required under free flow speed v_{ref} , taken to be 60 mph. So the delay in segment i in time t is

$$d_i(t) = l_i \times q_i(t) \times \left(\frac{1}{v_i(t)} - \frac{1}{v_{ref}} \right) \text{ vehicle-hours,} \quad (1)$$

in which l_i is the segment length in miles, and $v_{ref} = 60$ mph. The total daily delay in the freeway study

section is the delay over all segments and times,

$$D_{tot} = \sum_{i=1}^n \sum_{t=1}^T d_i(t). \quad (2)$$

Collisions are indexed $a = 1, 2, \dots$. The time t_a when a collision occurs and its location s_a are known, but the time it takes to clear the collision, and the spatial and temporal impact of the collision, are not known.

3.1 Decomposition Of Delay

The proposed method divides the total delay (1) into three components,

$$D_{tot} = D_{col} + D_{pot} + D_{rem}. \quad (3)$$

It will be useful to define

$$D_{rec} = D_{tot} - D_{col}. \quad (4)$$

Above

D_{col} is the total daily delay caused by collisions,

D_{rec} is the daily ‘recurrent’ delay,

D_{pot} is the potential reduction of D_{rec} by ramp metering, and

D_{rem} is the residual delay, most of which is attributed to excess demand.

D_{tot} , calculated from flow and speed data, is the additional vehicle-hours spent driving below the reference speed of 60 mph. D_{col} is the portion of this delay that is estimated to be caused by collisions. The difference, $D_{tot} - D_{col}$, is by convention (1, 7) called ‘recurrent delay’. A portion of the recurrent delay is due to frequently occurring bottlenecks, and could, in principle, be reduced by ramp metering. That potential reduction is estimated as D_{pot} . The remaining delay, D_{rem} , is due to all other causes, most of which is likely due to demand in excess of the maximum sustainable flow. The delay due to excess demand can only be reduced by changing trip patterns.

We now describe how each component is estimated.

3.2 Delay From Collisions

When a collision occurs, congestion propagates upstream of the collision location up to some maximum spatial *extent*. The congestion lasts a certain amount of time, called the collision’s *duration*. In dealing with empirical data, we declare a freeway segment i to be congested during a 5-minute time t if the speed $v_i(t) < 50$ mph. Using this definition of a congested state, we first identify the duration-extent ‘rectangle’ of a collision’s impact. We then calculate the total delay in this rectangle using formula (1).

To obtain the delay contribution of the collision by itself, however, we must subtract from this total delay the delay that would have occurred in the absence of the collision. We estimate this ‘recurrent’ delay D_{rec} by a nearest-neighbor prediction based on historical data.

This procedure is modified in the few cases in which the impact rectangle of one collision includes a second collision. The impact of the first collision is then considered to extend only up to the location of the second collision.

We present the procedure in detail below.

Identifying duration and extent of collisions

For each collision a , we first find the nearest segment i_a upstream of the (known) collision location s_a . We then check whether the speed in segment i_a is below 50 mph at any time within 15 minutes after t_a . This 15-minute ‘lag’ accommodates an error in the reported collision time t_a of up to 10 minutes and also allows some time for the impact of the collision to reach the detector in segment i_a .

If there is such a speed drop, we search for the longest consecutive time block $(t_a + 15 \text{ mins}, \dots, t_a + A_a)$ throughout which the speed at i_a is below 50 mph. This longest time block is collision a ’s duration.

In case the duration reaches another collision a' within the same segment i_a , at time say $t_a + A'_a \leq t_a + A_a$, we assign duration $(t_a + 15 \text{ mins}, \dots, t_a + A'_a)$ to collision a and start the process over for collision a' at time $t_a + A'_a$.

We identify the spatial extent of each collision a as follows. We first search for any other collision a' located upstream of s_a whose duration overlaps with the duration of collision a . If a' is such a collision, the extent of collision a is limited to the distance between the locations of the two collisions, $s_a, s_{a'}$. Within this bound, for each time $t \in (t_a, \dots, t_a + A_a)$, we search upstream until the speed recovers to above 50 mph to obtain the set of congested segments

$$B_a(t) = \{j < i_{a'} : v_k(t) < 50 \text{ mph, for all } k \text{ with } j \leq k \leq i_a\}.$$

The extent of collision a is the largest set of segments B_a among the $B_a(t)$, i.e.

$$B_a = \bigcup_{t=t_a}^{t_a+A_a} B_a(t).$$

Repeating this procedure for all collisions gives the duration-extent rectangle (A_a, B_a) , for each collision $a = 1, 2, \dots$.

Separating recurrent and non-recurrent congestion

Having identified the duration-extent of collision a , we compute its delay impact as follows. For each $t \in \{t_a, \dots, t_a + A_a\}$ in the duration, calculate the total delay at t within the spatial extent of the collision,

$$D_{tot,a}(t) = \sum_{i=i_a}^{i_a+B_a} d_i(t), \quad (5)$$

in which $d_i(t)$ is given in (1).

To obtain the delay at time t caused by the collision, we must subtract from $D_{tot,a}(t)$ the ‘recurrent’ congestion $D_{rec,a}(t)$ that would have occurred in the absence of the collision. We estimate this recurrent congestion as the K -nearest neighbor prediction of the recurrent delay, based on historical data of the delay $D_a(t, d)$ during the same time t and over the same spatial extent, for several other days $d = 1, \dots, D$. More precisely, the estimate of the recurrent congestion is the median value

$$D_{rec,a}(t) = \text{median}\{D_a(t, d'_k), k = 1, \dots, K\}, \quad (6)$$

in which $d'_k, k = 1, \dots, K$ are K days with smallest value $|D_a(t_a, d) - D_{tot,a}(t_a)|$ for $d = 1, \dots, D$. The recurrent congestion over the duration-extent of collision a is estimated to be

$$D_{rec,a} = \sum_{t=t_a}^{t_a+A_a} D_{rec,a}(t). \quad (7)$$

The extra delay caused by collision a is now given by

$$D_{col,a} = \sum_{t=t_a}^{t_a+A_a} \max(D_{tot,a}(t) - D_{rec,a}(t), 0), \quad (8)$$

and the estimate of the total delay caused by all collisions is

$$D_{col} = \sum_a D_{col,a}. \quad (9)$$

D_{col} is the collision ‘slice’ of the congestion pie (3). Because (8) gives the delay caused by each collision, we can obtain the frequency distribution (histogram) of collision-induced congestion. We can also correlate the collision-induced congestion with other traffic conditions such as volume or recurrent congestion (7). This detailed information can be useful in selecting effective mitigation strategies.

3.3 Potential Delay Reduction By Ramp Metering

Subtracting D_{col} from D_{tot} gives the recurrent delay D_{rec} (4), a large fraction of which is caused by frequently occurring bottlenecks. A portion of this bottleneck-induced delay can be removed by appropriate ramp metering. We describe the procedure to estimate this potential reduction, D_{pot} . The procedure involves two steps.

In the first step, we identify the frequently occurring bottlenecks, adapting an algorithm proposed in Chen et al. (12) (see also (13)). In the second step, we estimate the benefits from an ‘ideal’ ramp metering strategy at each bottleneck. This estimate is based on a procedure reported in (14) (see also (15)).

Bottleneck Identification

The algorithm of (12) is used to identify frequently occurring bottlenecks. The congested region associated with a bottleneck is a contiguous group of congested segments immediately upstream of the bottleneck location. The algorithm declares that there is an *active bottleneck* between two locations x_i and x_j , with $x_i < x_j$, at time t if the following inequalities hold:

- (1) $x_j - x_i < 2$ miles,
- (2) $v(x_k, t) - v(x_l, t) > 0$ if $x_i \leq x_k < x_l \leq x_j$
- (3) $v(x_j, t) - v(x_i, t) > 20$ mph, and
- (4) $v(x_i, t) < 40$ mph.

These inequalities essentially say that in the congested region, speed must decline monotonically as the bottleneck is approached from the upstream direction. The parameters in these inequalities (2 miles, 20 mph, 40 mph) are based on experimentation, as explained in (12). These inequalities are tested for each location of the freeway study section and each time t . In this way all active bottlenecks, together with their duration and extent, are identified.

We restrict attention to frequently occurring bottlenecks, defined to be those bottlenecks that are active for more than 20% of the days. The extent of each of these bottlenecks is taken to be the furthest upstream location that is congested for more than 20% of the days. This automatic identification of frequent bottlenecks, their duration and extent, only uses traffic data.

Additional information about the freeway geometry and traffic behavior should be used to modify the automatically identified bottleneck extent and duration and to ensure that the section is suitable for metering. This is the only manual step in our algorithm that requires human intervention, and may change the metering strategy described below. The manual step is of course not needed to apply the method.

Metering the bottlenecks

For each bottleneck, we calculate the potential reduction in delay by metering as follows. First, if for a particular day the bottleneck-related congested space-time region overlaps with the congestion impact region of a collision, that day is excluded from the computation.

For each remaining recurrent congestion-only day, let $x_i < x_j$ be the upstream and downstream boundaries of the bottleneck, respectively. Let $t_1 < t_2$ be the times one hour before the bottleneck is first activated and one hour after it disappears. From traffic data at detectors i and j we compute the cumulative volumes crossing the bottleneck boundaries,

$$Q_i(t) = \sum_{s=t_1}^t q_i(s), \quad Q_j(t) = \sum_{s=t_1}^t q_j(s),$$

beginning one hour before bottleneck activation at t_1 and ending one hour after its disappearance at t_2 . The difference between the total peak period volumes at the two locations,

$$\Delta Q = Q_j(t_2) - Q_i(t_2),$$

is the net inflow volume from on- and off-ramps between x_i and x_j during the duration of bottleneck activation.

This net inflow is responsible for activating the bottleneck. We assume that the bottleneck is caused by large volume surges at on-ramps so that $\Delta Q > 0$, and we meter the ramps to hold back these surges.

In some cases a bottleneck is created by a spillback from an off-ramp whose capacity cannot accommodate the exiting flow. In this case $\Delta Q < 0$. We do not consider such bottlenecks, which anyway cannot be eliminated by ramp metering.

To specify the metering strategy, we need the actual profiles of volume demand at the on-ramps. Unfortunately, ramp data in PeMS are incomplete and unreliable. So we resort to an approximation, based on two assumptions. First, assume that there is a single ‘virtual’ on-ramp between x_i and x_j located in section i with flow $\Delta q_i(t)$, so the total inflow into section i is

$$q'_i(t) = q_i(t) + \Delta q_i(t).$$

Second, assume that the time profile of $\Delta q_i(t)$ is proportional to the the average of the volume profiles at segment i and j . This gives the adjusted incoming volume into section i ,

$$q'_i(t) = q_i(t) + \Delta q_i(t) = q_i(t) + \frac{q_i(t) + q_j(t)}{Q_i(t_2) + Q_j(t_2)} \Delta Q.$$

This specification meets the consistency requirement that the cumulative flows are equal when the bottleneck disappears, i.e. $Q'_i(t_2) = Q_j(t_2)$.

We meter the adjusted flow q'_i . To set the metering rate, we first calculate from traffic data the ‘capacity’ C_j of the bottleneck section as the maximum flow at location j that is sustained for a 15-period interval. (For an estimate of these capacities for Los Angeles, see (16).) We meter the total inflow $q'_i(t)$ at 90% of C_j . Because the metered flow is always kept 10 percent below capacity, we assume that the resulting freeway traffic will be free flowing (60 mph). The only delay now experienced by vehicles is the queuing delay at the ramps, D_{rem} , which is given by the usual queuing formula,

$$D_{rem} = \sum_{t=t_1}^{t_2} \max \left(\sum_{s=t_1}^t q'_i(s) - 0.9C_j(t - t_1), 0 \right).$$

The difference, $D_{pot} = D_{rec} - D_{rem}$, is the portion of recurrent delay that can potentially be eliminated by appropriate ramp metering.

Empirically of course, D_{rem} includes the delay from all other causes such as lane closures, adverse weather, etc. We believe that a large part of D_{rem} is indeed the delay at the ramps caused by demand temporarily exceeding the maximum sustainable flow, and which cannot be further reduced by ramp metering.

3.4 Congestion Pie Chart

The method described above divides total delay D_{tot} into three parts: D_{col} , the collision-induced delay; D_{pot} , the delay that ramp metering could eliminate; and D_{rem} , the remainder, largely due to excess demand. Recall that the procedure does not calculate D_{pot} for days when a bottleneck-induced congested region overlaps a collision-induced congestion region. For those days, we assume that the ratio D_{pot} / D_{rec} is the same as for the days for which D_{pot} is computed.

Two types of pie charts are shown in Figure 6. One of these is the *congestion pie*—the division of congestion delay into three components; the other is the *VHT pie*—the division of all vehicle hours traveled (VHT) into free-flow VHT and congested VHT. The VHT pie gives an indication of ‘*congestion exposure*’, the fraction of travel time during which one may be exposed to congestion.

4 CASE STUDY

The method is applied to a 22.5 mile (postmile 4.5 to 27) section of northbound I-15 in San Diego County (Caltrans District 11). The time period is from 5AM to 10 PM, for 44 weekdays (September 2–October 31, 2002). There are 24 loop detectors in the section for which 5-minute lane-aggregated volume and speed data are obtained from the PeMS website (17). Data for collisions during this period are from Traffic Accident Surveillance and Analysis System (TASAS), maintained by the California Department of Transportation. (Data for 2002 rather than current data are used because there is an 18-month lag in the publication of TASAS data.)

Figure 1 shows the pattern of flow, speed, delay and accident risk by postmile and time of day. Clearly visible are a recurrent AM bottleneck near postmile 7 and a larger PM bottleneck near postmile 25. The average daily delay is 5,672 vehicle-hours. The delay contours match the speed contours much better than the flow contours. The crash rate is highest near the PM bottleneck. The crash rate is the number of collisions per hour per freeway mile per year. There were 74 collisions during the study period.

4.1 Delay From Collisions

Figure 2 illustrates how well the algorithm automatically delineates the space-time region of a collision impact. On October 3, 2002, there were three collisions, whose location and time of occurrence are indicated by their ID 1, 2, 3 on the top left speed contour plot. The three rectangles in the top right plot are the algorithm-generated impact regions of these collisions. The lower two plots show these regions for the seven collisions that occurred on October 11, 2002. Comparison with the speed contour plots suggest that the algorithm determines the space-time impact of collisions quite well. Having assigned the impact region for each collision a , formula (5) is used to calculate $D_{tot,a}(t)$.

The next step is to predict the recurrent congestion $D_{rec,a}(t)$ for each collision using the K -nearest neighbors in formula (6). Figure 3 illustrates this calculation for collisions #2 and #5 on October 11, 2002. Each figure shows 44 delay profiles (gray lines), one for each day. Each gray line is a plot of the 5-minute delay incurred at that location from 5AM to 10PM. The dark bold profile is for October 11, the day of the

collisions. These data are from PeMS.

The first vertical line is drawn at the time of the collision, obtained from TASAS data. The second vertical line marks the extent of the collision impact, obtained from the calculations illustrated in Figure 2. The bold dotted line is the predicted delay, calculated from formula (6).

For collision #2, the total delay is 1,226 vehicle-hours, and the delay just before the collision is zero. For the same segment, most other days that exhibit zero delay at the time of collision #2, continue to have very small delay for the rest of the day, and this leads to a prediction of almost zero delay. The total predicted delay over the duration is thus zero and all 1,226 vehicle-hours are assigned to the collision-induced delay.

By contrast, for collision #5, the total delay is very large, 5,195 vehicle-hours, but nearly 80% of this delay (4,110 vehicle-hours) is predicted to be recurrent: This delay would have occurred even in the absence of this collision. Consequently, only 1,085 vehicle-hours is attributed to collision #5.

Figure 4 exhibits the distribution of delays incurred by the accidents. On average, each collision induces a delay of 477 vehicle-hours. Only 25 of 74 accidents (33%) cause any delay and nearly 70% of collisions cause no delay. The distribution illustrates the '10-90 rule': 10 percent of collisions account for 90 percent of collision-induced delay.

The 477 vehicle-hours of delay per collision is roughly in agreement with other estimates. A regression of total daily delay vs. number of accidents for all of Los Angeles yields a slope of 560 vehicle-hours per accident, see (2, p.20). For southbound I-5 in Seattle, Hallenbach et al. find that a lane-blocking incident causes between 318 (conservative estimate) and 591 (liberal estimate) vehicle-hours of delay (1, p.15).

The average daily delay caused by collisions, D_{col} , is 802 vehicle-hours, which is 12.4% of total daily delay. By way of comparison, Hallenbach et al. find that "for the urban freeways examined [in the Central Puget Sound region of Washington State] lane-blocking incidents are responsible for between 2 and 20 percent of total daily delay (1, p.8). These average numbers must be used with caution because the delay impact of collisions varies considerably from freeway to freeway and over different times of day. For example, in our case study, during the AM peak (5 AM to 10 AM), the average daily delay and average collision-induced delay are only 393 and 20 (or 4.9%) vehicle-hours.

Figure 5 further clarifies this issue. It shows that collisions that occur in the morning cause little delay. In the afternoon, when there is high recurrent congestion, there is a greater chance of collision and greater delay, especially if the collision occurs at the beginning of the recurrent congestion. Plots like Figure 5 can help design the freeway locations and times during which to deploy accident management resources like the Freeway Service Patrol (FSP), see (18).

Another note of caution is introduced by recognizing that TASAS data are incomplete and the fact that the case study does not consider non-collision incidents. Of course if collisions are randomly omitted by TASAS, the distribution of delay per incident obtained above is not affected. But the total contribution of collision can change. We compare TASAS collision counts with a separate, supposedly more detailed incident data source from FSP studies (4, 5). For fiscal year 2002-2003, the total number of incidents in the study section is 11,380, of which 8.17% or 929 are accidents and the remaining 92% are breakdowns and debris. By comparison, TASAS reports 430 crashes, or 46% of accidents in the FSP study.

4.2 Ramp Metering And The Pie Chart

The algorithm automatically identifies two frequent bottleneck activations at the following locations and time periods:

- Location 5 at postmile 7.581, 6:45 AM - 8:25 AM,
- Location 19 at postmile 24.511, 5:40 PM - 6:00 PM.

We calculate the potential reduction in delay resulting from ramp metering at these two bottlenecks. We meter the 4.2-mile section determined by detectors VDS 1108588 (postmile 19.67) and 1108595 (postmile 23.867) for the PM period and the 2-mile section determined by detectors VDS 1108676 (postmile 6.381) and 1108528 (postmile 7.2) for the AM period.

The AM and PM peak periods experience 39 (88%) and 22 (50%) collision-free days. For those days, the original delay (D_{rec}) of 2,863 and 271 vehicle-hours is reduced by ramp metering, to 432 and 66 vehicle-hours, or a reduction of 85% and 76%. Assuming that proportionately the same reduction can be achieved for the days with collision, the average potential reduction in delay is 2,636 vehicle-hours.

The total daily delay caused by collisions, 'non-ideal' metering, and the remainder, attributed to excess demand, is then computed to be 1,733, 2,636 and 1,315 vehicle-hours or 31%, 46% and 23%. Figure 6 shows the congestion and VHT pie charts.

4.3 Relationship Between Collision-Induced And Recurrent Delay

The method produces the distributions of duration, extent, and delay caused by individual collisions. This information can be correlated with typical volume and delay at the section near the collision to investigate the relationship (if any) between collision-induced and recurrent congestion. Figure 7 shows scatter plots of collision-induced delay (on the y axis) versus historical volume, the K -nearest neighbor predicted recurrent delay, and historical (total) delay.

Each plot also displays a nonparametric regression curve fitted through the scatter plot. In each case, the regression curve has a parabolic shape, implying that the average impact of collision on congestion is most severe when the freeway is moderately congested with high volume: Collision-induced delay decreases as the historical volume and delay get very small or very large. This appears intuitive, and the regression curves lend empirical confirmation. The empirical relationship may be used to better target incident-management strategies such as the FSP.

5 CONCLUSION

Between 1980 and 1999, route miles of highways increased 1.5 percent while vehicle miles of travel increased 76 percent, according to FHWA (11). In 2000, the 75 largest metropolitan areas experienced 3.6 billion hours of delay, resulting in \$67.5 billion in lost productivity, according to the Texas Transportation Institute. Congestion mitigation is high on the agenda of transportation agencies, and the key to this is a better understanding of the causes of congestion and their relative impact, summarized in the pie of congestion.

As more comprehensive data become available, and more powerful statistical methods are used to analyze these data, more reliable quantitative estimates of the impact of different causes will be obtained. Reliability of the estimates will improve as systematic methods are applied to large data sets from different freeways. For such a program of analysis to be realized, these methods have to be automated.

The paper proposes a fully automated method that requires traffic data (volume or flow and speed) and data on the location and time of occurrence of collisions. With these data, the method divides the total delay into three components: delay attributed to collisions, delay that can be eliminated by ramp metering, and the remaining delay due to all other causes, but which we believe is due mainly to excess demand.

The calculation of the delay caused by a collision of course requires predicting the (recurrent) delay that would have occurred in the absence of that collision. The proposed method first delineates the space-time region of the collision impact and then uses a K -nearest neighbor estimate to predict the recurrent delay in the impact region.

The calculation of how much delay can be reduced by ramp metering requires first the automatic location of frequently-occurring bottlenecks, and then an estimate of the benefits from an appropriate ramp metering strategy.

The method produces estimates of the distribution of spatial extent, duration, and delay impact of collisions. That information can be used to better target mitigation strategies.

The method is applied to a 24-mile section of northbound I-15 in San Diego. It estimates an average delay of 477 vehicle-hours per collision, which is within the ranges obtained in other studies, based on very different analyses. All collisions taken together account for 12.4% of total delay.

The method also finds that 46% of total delay could, in principle, be eliminated by appropriate ramp metering. This is a surprisingly large fraction, for which there is some independent but indirect support. If more extensive studies confirm such a large estimate, it suggests that very large productivity improvements in freeway operations are within reach.

ACKNOWLEDGEMENT

We are grateful for comments and criticism from John Wolf and Fred Dial of Caltrans; Tarek Hatata of the System Metrics Group; and Chao Chen, Alex Skabardonis and Karl Petty of the PeMS Development Group.

This study is part of the PeMS project, which is supported by grants from Caltrans to the California PATH Program. The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views of or policy of the California Department of Transportation. This paper does not constitute a standard, specification or regulation.

REFERENCES

- [1] M.E. Hallenbach, J.M. Ishimaru, and J. Nee. Measurement of recurring versus non-recurring congestion. Washington State Transportation Center (TRAC), October 2003.
- [2] System Metrics Group. Freeway performance report. Prepared for California Department of Transportation, 2003.
- [3] California Department of Transportation. 2002 HICOMP Report. State Highway Congestion Monitoring Program, November 2003.
- [4] K. Petty, H. Noeimi, K. Sanwal, D. Rydzewski, A. Skabardonis, P. Varaiya, and H. Al-Deek. The freeway service patrol evaluation project: Database support programs, and accessibility. *Transportation Research, Part C*, 4(2):71–85, April 1996.
- [5] A. Skabardonis, K. Petty, P. Varaiya, and R. Bertini. Evaluation of the freeway service patrol (fsp) in los angeles. Research Report UCB-ITS-PRR-98-31, California PATH, University of California, Berkeley, CA 94720, 1998.
- [6] R. Bertini, S. Tantiyanugulchai, E. Anderson, R. Lindgren, and M. Leal. Evaluation of Region 2 Incident Response Program using archived data. Transportation Research Group, Portland State University, July 2001.
- [7] A. Skabardonis, K. Petty, and P. Varaiya. Measuring recurrent and non-recurrent traffic congestion. In *Proceedings of 82nd Transportation Research Board Annual Meeting*, Washington, D.C., January 2003.

- [8] D. Bremmer, K.C. Cotton, D. Cotey, C.E. Prestrud, and G. Westby. Measuring congestion: Learning from operational data. In *Proceedings of 83rd Transportation Research Board Annual Meeting*, Washington, D.C., January 2004.
- [9] Transportation Research Board. *Highway Capacity Manual 2000*, December 2000.
- [10] S.M. Chin, O. Franzese, D.L. Greene, H.L. Hwang, and R.C. Gibson. Temporary losses of highway capacity and impacts on performance. Technical Report ORNL/TM-2002/3, Oak Ridge National Laboratory, National Transportation Research Center, Knoxville, TN 37932, May 2002.
- [11] FHWA Congestion Mitigation website. <http://www.fhwa.dot.gov/congestion/congest2.htm>.
- [12] C. Chen, A. Skabardonis, and P. Varaiya. Systematic identification of freeway bottlenecks. In *Proceedings of 83rd Transportation Research Board Annual Meeting*, Washington, D.C., January 2004.
- [13] L. Zhang and D. Levinson. Some properties of flows at freeway bottlenecks. In *Proceedings of 83rd Transportation Research Board Annual Meeting*, Washington, D.C., January 2004.
- [14] Z. Jia, P. Varaiya, C. Chen, K. Petty, and A. Skabardonis. Congestion, excess demand and effective capacity in California freeways. Online at pems.eecs.berkeley.edu, December 2000.
- [15] C. Chen, Z. Jia, and P. Varaiya. Causes and cures of highway congestion. *IEEE Control Systems Magazine*, 21(4):26–33, December 2001.
- [16] Z. Jia, P. Varaiya, C. Chen, K. Petty, and A. Skabardonis. Maximum throughput in LA freeways occurs at 60 mph v. 4. Online at pems.eecs.berkeley.edu, January 2001.
- [17] PeMS website. <http://pems.eecs.berkeley.edu>.
- [18] K. Petty, M. Ostland, J. Kwon, J. Rice, and P. Bickel. A new methodology for evaluating incident detection algorithms. *Transportation Research, Part C*, 10(3):189–204, June 2002.

LIST OF FIGURES

- 1 Space and time-of-day distribution of flow (veh/5-minute), speed (mph) and delay (veh-hour) in the northbound I-15, averaged over 44 weekdays. Vehicles travel from the bottom of the plot to the top in increasing postmile. The last plot shows the crash rate (accidents/hour/mile/year) estimated by density estimation, and the time and location of individual collisions in dots. 14
- 2 Application of algorithm delineating the space-time region affected by the collisions of October 3, 2002 (Top) and October 11, 2002 (Bottom). The plots on the left show the space-time distribution of speed for the day with collision time and location marked with an ID. The plots on the right shows, in the same space-time axis, the regions affected by each collision. 15
- 3 Observed and K -nearest neighbor predicted delay profile associated with collisions #2 (top) and #5 (bottom) on October 11, 2002. The observed (solid black line) and predicted (dotted black line) delay profiles are shown as well as the delay profiles of 43 other days (light gray lines) used for prediction. Two vertical lines are the previously estimated boundaries of the collision duration. The observed total delay is 1,226 and 5,195 vehicle-hours for the two collisions and the predicted delay is 0 and 4,110 vehicle-hours. 16
- 4 Distribution of delay caused by collisions. Out of 74 collisions, 49 collisions cause no delay at all and only 25 cause any delay. The average delay per collision is 477 vehicle-hours. . . . 17
- 5 The observed, predicted (recurrent) and collision delay associated with each collision and their relationship over space and time. Individual crashes are shown by 'x' and the area of the circles is proportional to the delay. Collision delay is the observed delay minus the predicted delay. 18
- 6 The congestion pie (left) and the VHT pie (right) automatically constructed for I-15 study section. 19
- 7 Relationship between Collision Delay and Historical Volume, Recurrent Delay, Total Delay. For standardized comparison, duration and extent of all collision are now fixed at 1.5 hours and 5 miles. Individual point in each plot corresponds to one of 74 collisions. The curve was obtained using local polynomial regression fitting with symmetric error distribution, degree 2 and span parameter 0.75. For the first two plots, excess delay is transformed by $y = x^{1/4}$ to highlight the functional relationship. 20

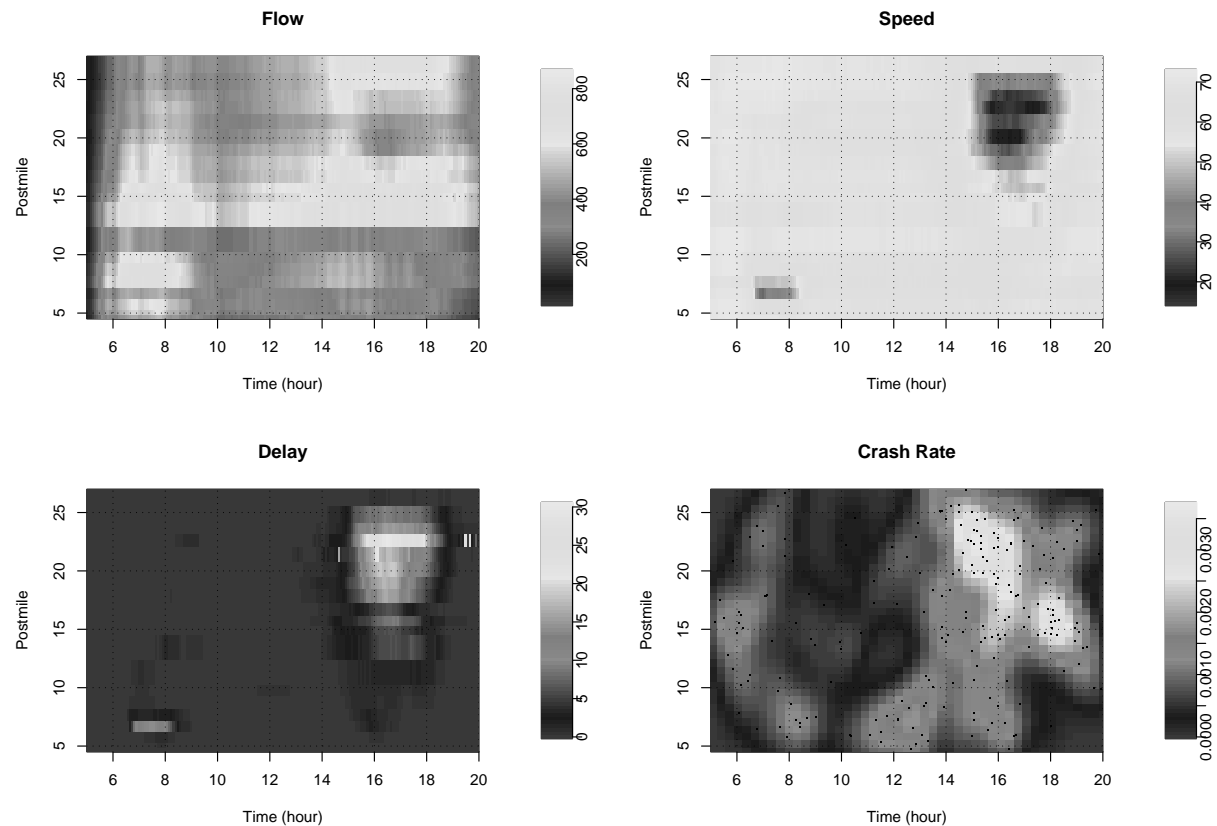


FIGURE 1 Space and time-of-day distribution of flow (veh/5-minute), speed (mph) and delay (veh-hour) in the northbound I-15, averaged over 44 weekdays. Vehicles travel from the bottom of the plot to the top in increasing postmile. The last plot shows the crash rate (accidents/hour/mile/year) estimated by density estimation, and the time and location of individual collisions in dots.

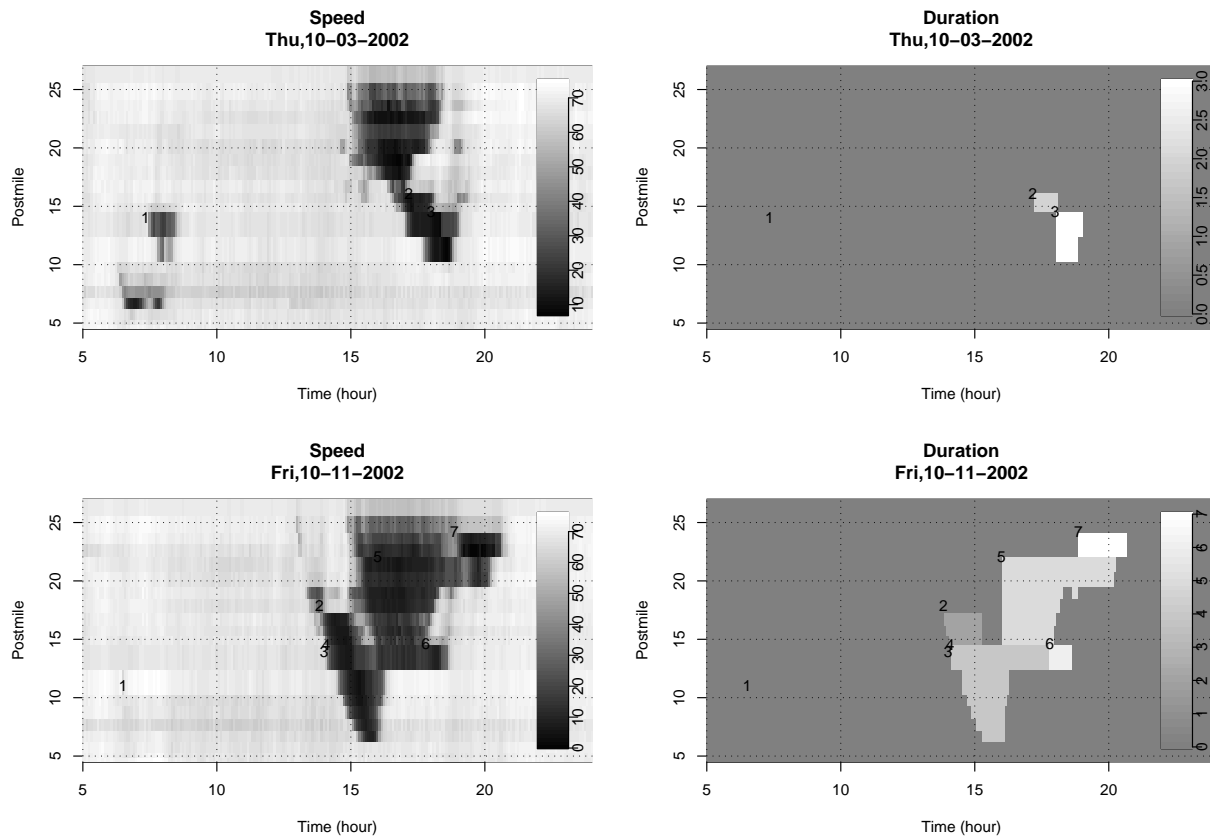


FIGURE 2 Application of algorithm delineating the space-time region affected by the collisions of October 3, 2002 (Top) and October 11, 2002 (Bottom). The plots on the left show the space-time distribution of speed for the day with collision time and location marked with an ID. The plots on the right shows, in the same space-time axis, the regions affected by each collision.

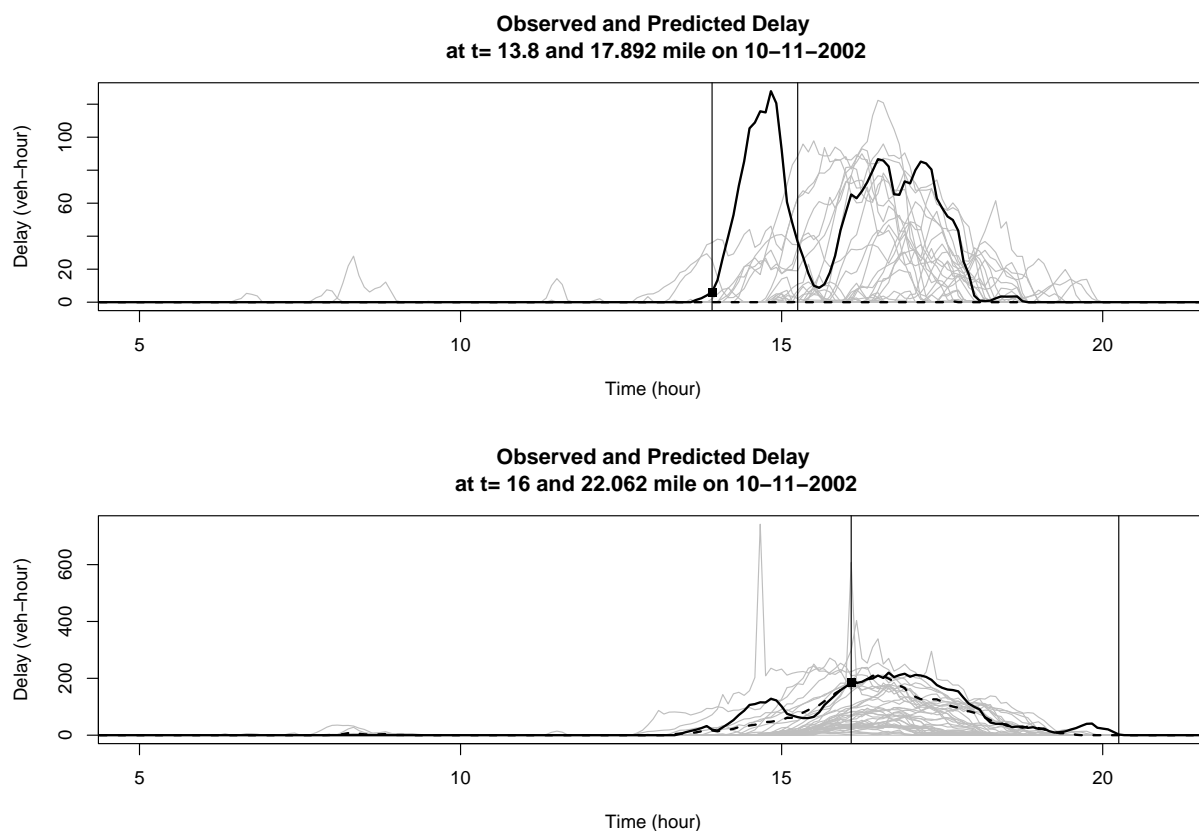


FIGURE 3 Observed and K -nearest neighbor predicted delay profile associated with collisions #2 (top) and #5 (bottom) on October 11, 2002. The observed (solid black line) and predicted (dotted black line) delay profiles are shown as well as the delay profiles of 43 other days (light gray lines) used for prediction. Two vertical lines are the previously estimated boundaries of the collision duration. The observed total delay is 1,226 and 5,195 vehicle-hours for the two collisions and the predicted delay is 0 and 4,110 vehicle-hours.

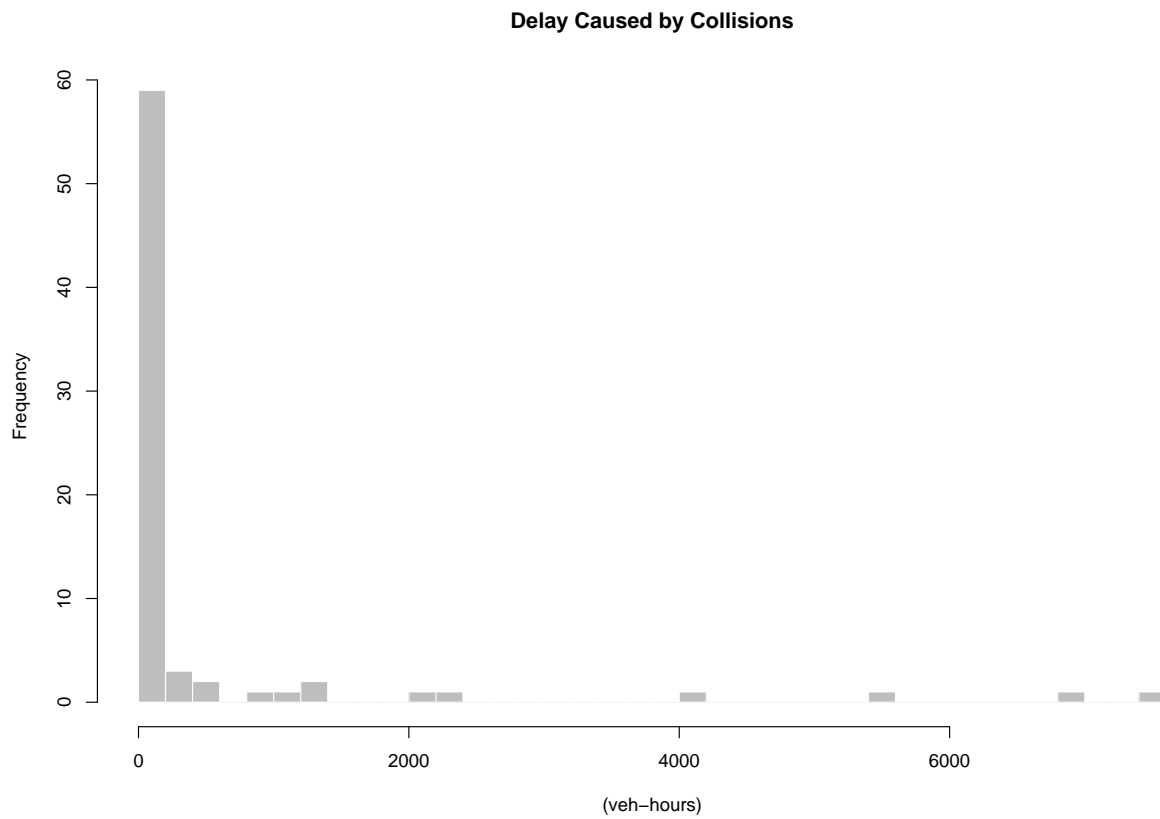


FIGURE 4 Distribution of delay caused by collisions. Out of 74 collisions, 49 collisions cause no delay at all and only 25 cause any delay. The average delay per collision is 477 vehicle-hours.

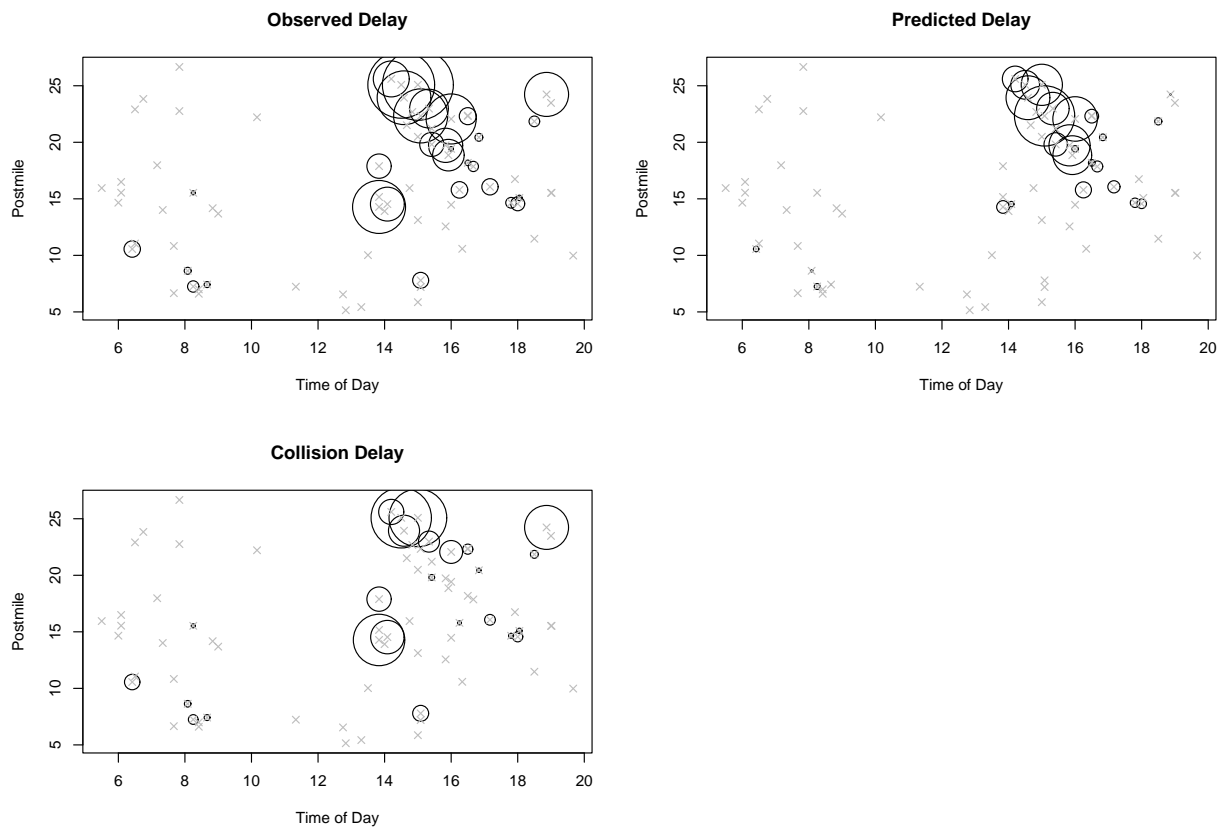


FIGURE 5 The observed, predicted (recurrent) and collision delay associated with each collision and their relationship over space and time. Individual crashes are shown by 'x' and the area of the circles is proportional to the delay. Collision delay is the observed delay minus the predicted delay.

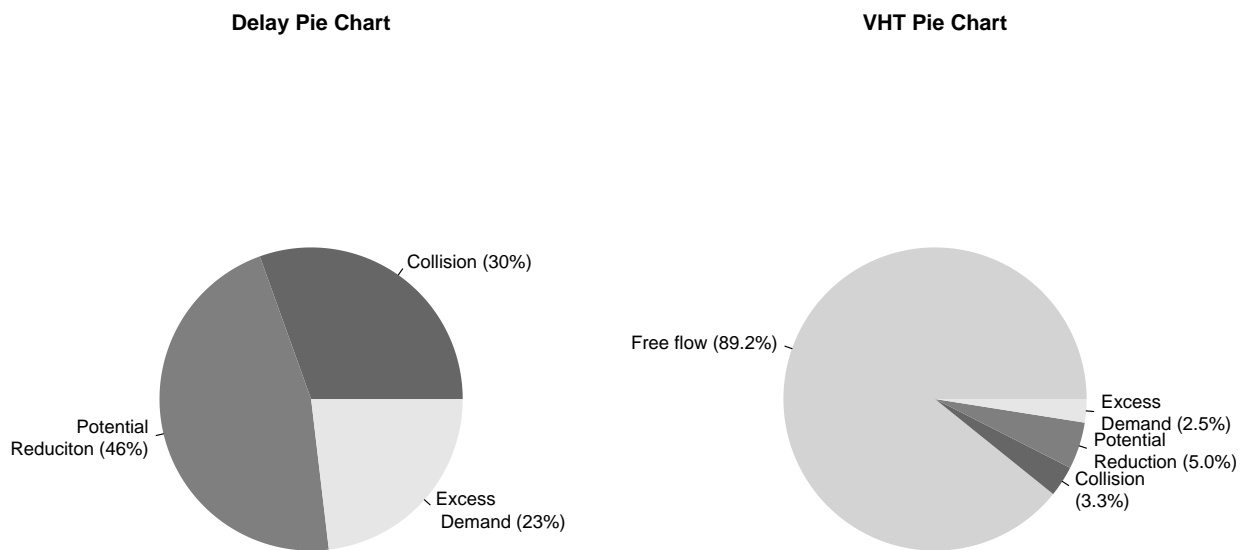


FIGURE 6 The congestion pie (left) and the VHT pie (right) automatically constructed for I-15 study section.

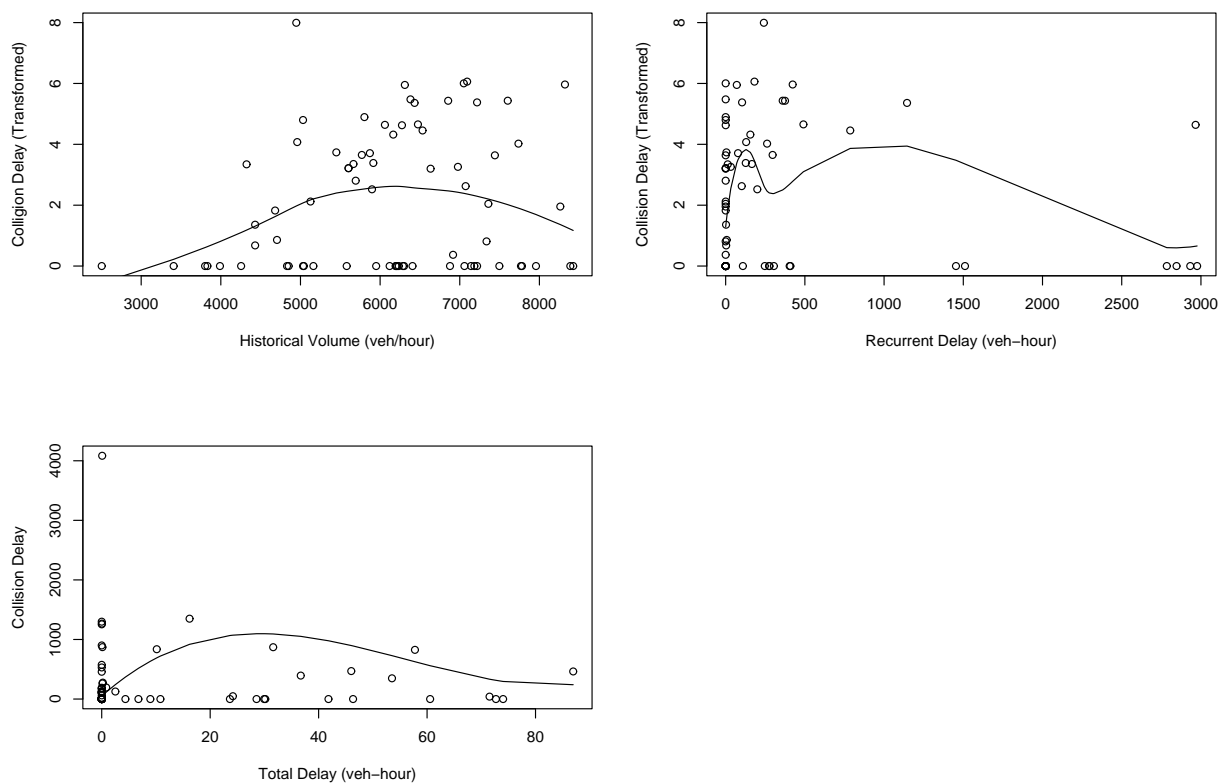


FIGURE 7 Relationship between Collision Delay and Historical Volume, Recurrent Delay, Total Delay. For standardized comparison, duration and extent of all collision are now fixed at 1.5 hours and 5 miles. Individual point in each plot corresponds to one of 74 collisions. The curve was obtained using local polynomial regression fitting with symmetric error distribution, degree 2 and span parameter 0.75. For the first two plots, excess delay is transformed by $y = x^{1/4}$ to highlight the functional relationship.