Relating Traffic Events to CHP Incidents
ECI 256 Final Project

Clark Fitzgerald, Angela Tobin

December 2016

Abstract

Image processing techniques were used to detect regions of unusally high vehicle occupancy within PeMS data for I80 West in the month of April 2016. These high occupancy events were then related to CHP incident reports.

1 Introduction

What

A traffic event is

2 Review

TODO: This needs to be improved

[Chung et al., 2007] describe traffic characteristics at three fixed bottlenecks.

[Chen et al., 2004] describe systematic identification of bottlenecks from 5 minute loop data. We want to do a similar thing with 30 second data on a larger scale. They use velocity measurements, which are not always reliable / available. So it might be better in our case to use occupancy and flow.

The say: "Five-minute data provide sufficient resolution for this analysis because the traffic features sought are on the order of 30 min or more." So maybe we look for finer features.

[Hall and Agyemang-Duah, 1991] use flow and occupancy since velocity is not available.

> Averages of flow and occupancy across the three lanes were used. The two tended to vary together in the period before congestion aud diverge during the congested period. Determining the exact beginning and end of congestion was, however, difficult from these numbers, so the ratio of occupancy to flow was used. Three

> values of the ratio were tested for the threshold level: 1.0, 1.1, and 1.2. A ratio of 1.0 gives a longer duration of bottleneck flows, some of which were very low, suggesting that demand was below capacity. A ratio of 1.2 excludes sustained periods (10 min) of high flows (5,800 vehicles/hr or more). A ratio of 1.1 or above persisting for 3 min was selected as the criterion for the identification of the start of a queue.

[Wieczorek et al., 2010] applies this to Oregon data.

[Zhang and Levinson, 2004] show that Queue Discharge Flows QDF's, normal around 2K passenger cars per lane per hour.

TODO: find papers that quantify traffic impacts of construction and various types of accidents.

3 Data Preparation

5 minute observation data for weekdays in April was downloaded in bulk from California's PeMS system. April was chosen because it's the first month of the year without holidays. Weekdays were used to avoid less regular traffic patterns on the weekend. The raw 30 second observations were also tried, but these were excessively variable and noisy, which makes them less suitable for this analysis.

PeMS defines the variable occupancy used as "Average occupancy across all lanes over the 5-minute period expressed as a decimal number between 0 and 1." Taking the average over time and all lanes is useful in reducing the variance of the occupancy, which is why it was used over the occupancy in one particular lane. In [Daganzo and Daganzo, 1997] Daganzo justifies the use of occupancy as a proxy for congestion.

Let x_{ijk} be the occupancy value on the ith day, tth 5 minute time interval, and mth mile marker. Let $\bar{x}_{\cdot jk}$ be the median value for across all 21 weekdays in April. A derived variable was formed by taking the difference

$$y_{ijk} \equiv x_{ijk} - \bar{x}_{\cdot x_{ijk}} \tag{1}$$

All further analysis centered on these differences.

Figure 3 shows the standard deviations of the difference y_{ijk}. The bright line around mile 8 marks the toll plaza to enter San Francisco. Lighter regions occur during the morning rush hour starting just before 7 AM, and all along the area between mile markers 0 and 15 during the day time, which corresponds to the area of high traffic between San Francisco and Berkeley. This shows that occupancy exhibits significant variablility in regions of congested traffic. The implication for this analysis is that the difference y_{ijk} will have more noise in these regions, which makes it more difficult to accurately detect traffic events of high occupancy.

Figure 3 displays the differences y_{ijk} as defined in equation 1 on April 25th. Corresponding CHP incidents have been plotted in the same graph. Two areas of high occupancy exist

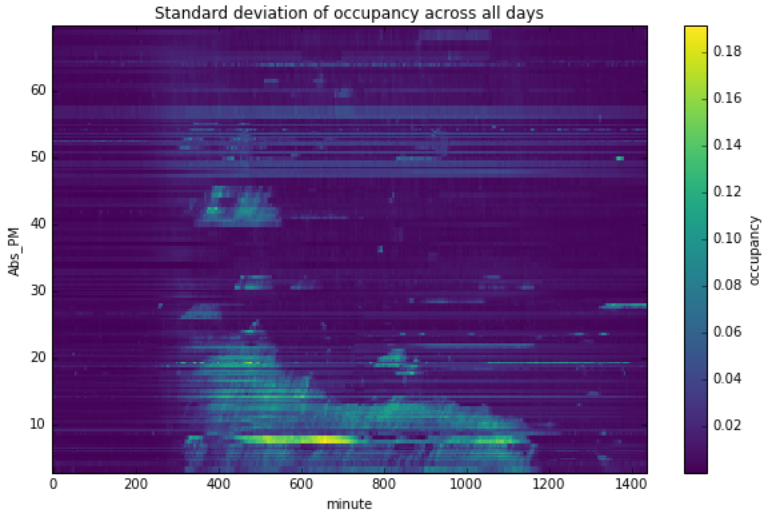

Figure 1: Occupancy varies more in areas of high traffic.

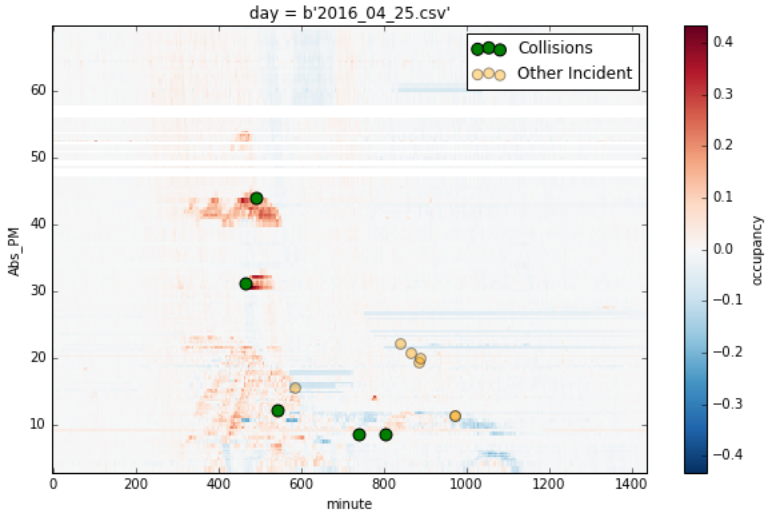

Figure 2: This shows the difference in occupancy from the median. Areas of unusually high occupancy are colored dark red.

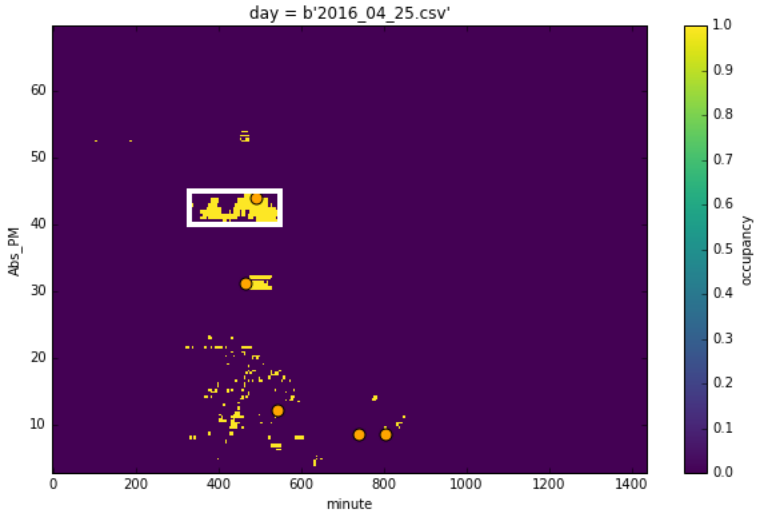

Figure 3: Areas of high occupancy have been converted to a binary variable.

around minute 400, one at mile 30 and one just above mile 40. Since there are green points representing collisions in these regions it seems reasonable to associate the traffic event with CHP incident data.

Simple image processing techniques were used to isolate and quantify these areas of high occupancy. The first step was to use simple thresholding to create a new binary variable that flags every observation that's larger than a certain size. Mathematically, let $z_{ijk} = 1$ if $y_{ijk} > t$ and $z_{ijk} = 0$ if $y_{ijk} \leq t$. Some experimentation showed $t = 0.1$ to be a reasonable threshold value. This produces figure 3. This resulting variable z_{ijk} was then treated as an image; shapes were inferred by finding bounding boxes for connected components as in figure 3. These bounding boxes then define what we consider to be a traffic event.

Detected traffic events have been plotted with CHP incident reports in figure 3. Most events occur during the morning rush and in the region between miles 0 and 15, which is the region between Berkeley and San Francisco. So events are concentrated in areas of high traffic.

4 Statistical Analysis

CHP incident reports were linked to bounding boxes if they occurred either within a bounding box, or slightly outside of one. This was done because a CHP event might be reported immediately before traffic conditions deteriorate. A tolerance of 1 mile and 10 minutes was chosen since mile resolution of the loop detectors is usually under 1 mile, and time resolution is 5 minutes. In figure 4 we see that most events could not be linked. This is unsurprising, since not all CHP events will be associated with a major traffic disturbance. An example of this is the CHP responding to a call on smuggled fishing boats.

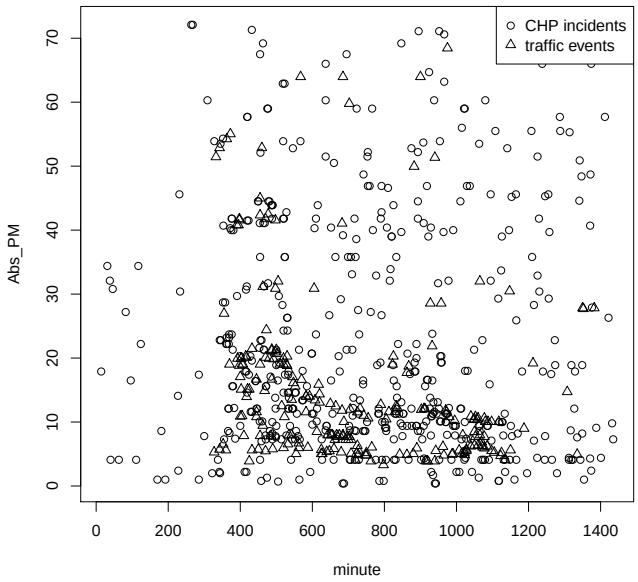

Figure 4: Events are concentrated in regions of heavy traffic and congestion.

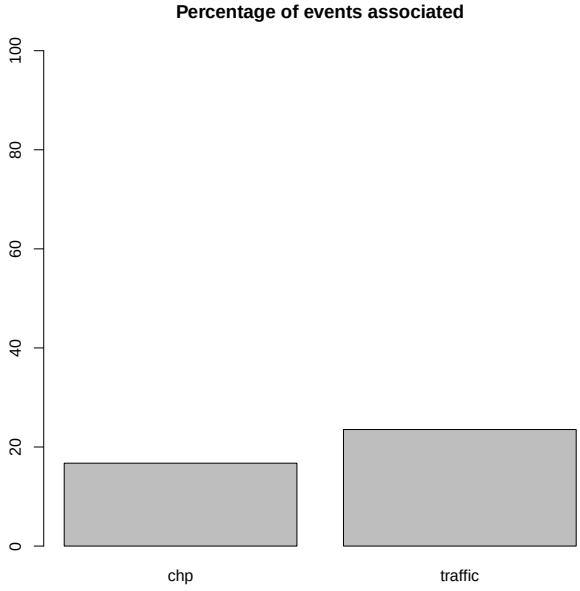

Figure 5: Most events could not be linked.

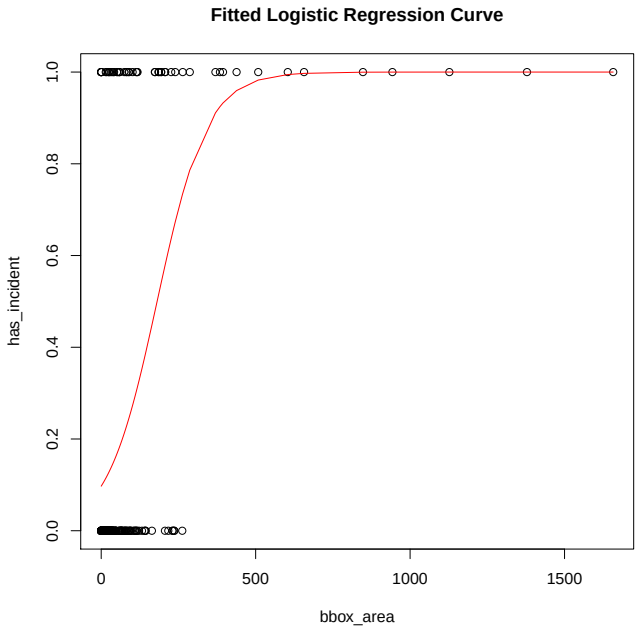

Figure 6:

Not all traffic events had associated CHP incidents, however those with large bounding boxes always did. Traffic events impacting an area of 2 miles for 2 hours correspond to a bounding box of 240 units. A logistic regression model on this data showed that such an event had a 90% chance of being associated with a CHP incident. Let $y = 1$ if an area of high occupancy has a CHP incident associated with it, and $y = 0$ otherwise. Let x be the bounding box area. The logistic regression model as shown in figure 4 is

$$\mathrm{E}(y|x) = \frac{e^{b_0+b_1 x}}{1 + e^{b_0+b_1 x}}.$$

A 95% confidence interval for b_1 is (0.008, 0.017). Since this region is positive we conclude that larger bounding box areas are more likely to have associated CHP incidents. We have to be a little careful with interpreting this however, since if a bounding box is large enough then it will include CHP incidents with which it's not actually related.

We did not examine any causal relationships between CHP incidents and traffic conditions, but this would be a very interesting line of further inquiry. Thinking about a vehicle collision as a CHP incident it's possible that CHP incidents cause traffic or that traffic causes CHP incidents. Both seem plausible.

A linear regression model was fitted using 13 CHP incident types as a predictor for the area of the bounding box. CHP incidents types include `type1125-Traffic Hazard, type1182-Trfc Collision-No Inj, type20001-Hit and Run w/Injuries`, etc. With a p value of 0.99 this model explained less than 1% of the variance in the area of the bounding box. Hence we found no evidence here of a relationship between the type of incident and the severity of the traffic incident. This does not mean that such a relationship does not exist; it just means

that this particular predictor in this data doesn't capture it. For example, it's common knowledge that a high speed crash blocking multiple lanes will lead to congested traffic, but this isn't necessarily captured in the incident type.

5 Discussion

An interesting further analysis would be to examine the actual shapes of the traffic events rather than bounding boxes. For example, given sufficient resolution a temporary bottleneck placed in uniform traffic and then removed should result in an inverted triangular shape.

Refining this analysis and scaling it up to a a larger scale, say all roads in the Bay Area for multiple years would provide interesting opportunities to detect and examine long term trends.

TODO:

References

[Chen et al., 2004] Chen, C., Skabardonis, A., and Varaiya, P. (2004). Systematic identification of freeway bottlenecks. *Transportation Research Record: Journal of the Transportation Research Board*, (1867):46–52.

[Chung et al., 2007] Chung, K., Rudjanakanoknad, J., and Cassidy, M. J. (2007). Relation between traffic density and capacity drop at three freeway bottlenecks. *Transportation Research Part B: Methodological*, 41(1):82–95.

[Daganzo and Daganzo, 1997] Daganzo, C. and Daganzo, C. (1997). *Fundamentals of transportation and traffic operations*, volume 30. Pergamon Oxford.

[Hall and Agyemang-Duah, 1991] Hall, F. L. and Agyemang-Duah, K. (1991). Freeway capacity drop and the definition of capacity. *Transportation research record*, (1320).

[Wieczorek et al., 2010] Wieczorek, J., Fernández-Moctezuma, R., and Bertini, R. (2010). Techniques for validating an automatic bottleneck detection tool using archived freeway sensor data. *Transportation Research Record: Journal of the Transportation Research Board*, (2160):87–95.

[Zhang and Levinson, 2004] Zhang, L. and Levinson, D. (2004). Some properties of flows at freeway bottlenecks. *Transportation Research Record: Journal of the Transportation Research Board*, (1883):122–131.

Maybe include?

Occupancy data was treated as an image. To compute the Once we have the denoised difference we can do the following:

1. (Optional) Detect shapes that are flat on top, since this is the distinguishing feature of a bottleneck.
2. Compute centroid, bounding boxes, and area which will quantify the impact in terms of space and time.
3. Join these features to incident data. This likely will require some text analysis.

Then we can answer questions such as:

1. How many traffic events occur which have no associated incident data? And what sort of events were they probably?
2. What is the impact of an event of type X on a given section of highway? Something along the lines of: When traffic flow is 1500 veh per lane per hour in a two lane freeway a collision involving exactly two vehicles typically creates congestion lasting 10-15 minutes which propagates back 2-3 miles.
3. How can we model the distribution of traffic incidents, ie. Poisson with some parameters.

But how can these results be useful more broadly? More accurate simulations. Input to real time routing. Impacts of planned construction events. Scheduling CHP patrols and recovery services.