Relating Traffic Events to CHP Incidents
ECI 256 Final Project

Clark Fitzgerald

December 2016

Abstract

In this project 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

[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.

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 around minute 400, one at mile 30 and one just above mile 40. Since there are green points

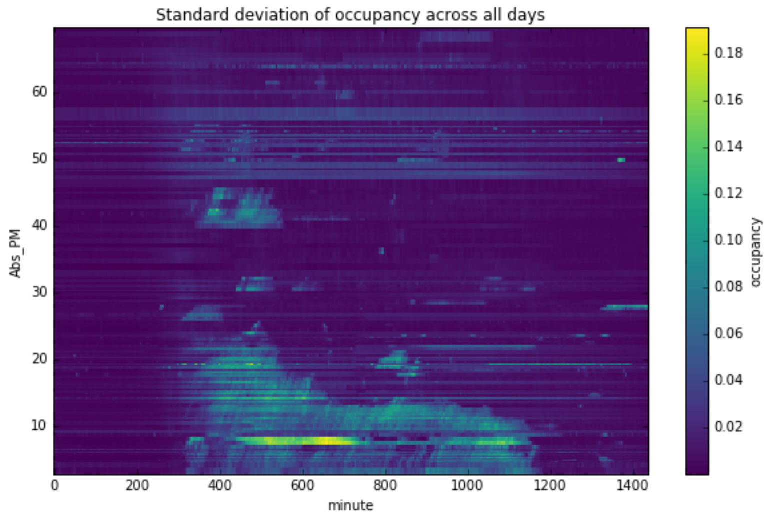

Figure 1: Occupancy varies more in areas of high traffic.

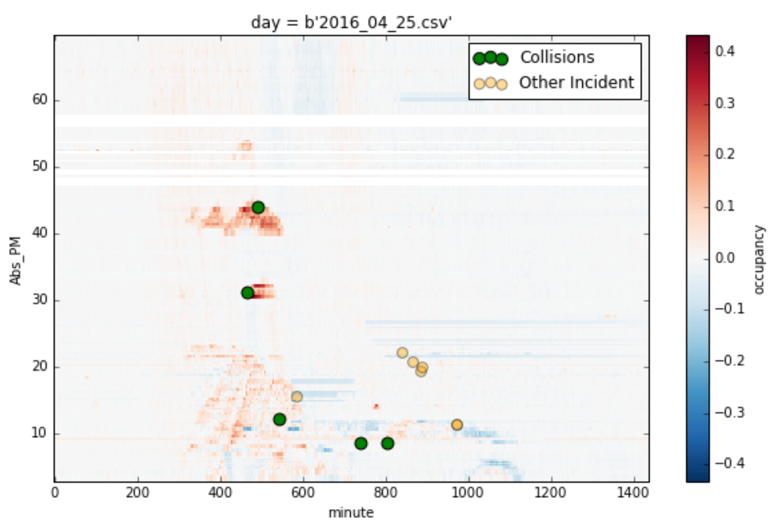

Figure 2: This shows the difference in occupancy from the median. Areas of unusually high occupancy are colored dark red.

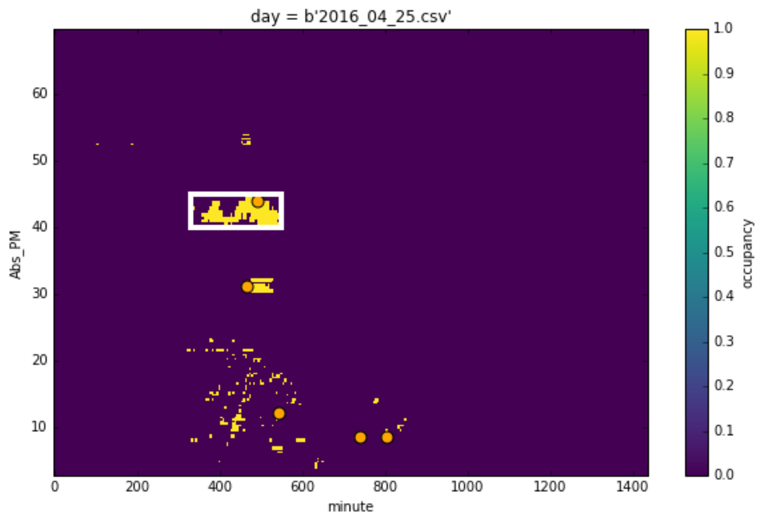

Figure 3: Areas of high occupancy have been converted to a binary variable.

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.

Occupancy data was treated as an image. To compute the

We can use opencv to do this. For each freeway we can take the following steps:

1. Read the raw file and convert it to an image with many missing values.
2. Interpolate missing values.
3. Threshold the image, so for a pixel with density $\rho > \rho_0$ we define it as high density, otherwise low density. We'll have to experiment to see what the appropriate value of ρ_0 is but I suspect around 0.3.
4. Once this has been done for every day we can compute an average thresholded image showing the areas of high density. This will show recurring bottlenecks. May need to threshold this also.
5. Compare each day with the average. The difference will show the unusual patterns that occurred on just one day. Might have to do some denoising here- eliminate points that are not part of a cluster.

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.

4 Statistical Analysis

About

5 Conclusions

My favorite broad idea: Empirically quantify the impact of various flow levels on congestion. I suspect that for a given type of accident total delay experienced by travelers increases superlinearly as a function of (flow / capacity). Assuming we're looking at uncongested regions. If this holds then it can be used as an argument that we need to try to keep (flow / capacity) under some level. This could mean taking measures such as toll roads or tax credits for employers who find ways to keep employees off roads during peak commute hours. My favorite- an argument to make cars more expensive (like tolls and mileage tax) and use money to improve public transportation.

Why is this interesting? Interdisciplinary since we combine data munging, computer vision, and statistics. Provides motivation for developing automatic R / C++ bindings that I'll be working on. Has legitimate possible impacts. Most computer vision stuff in traffic works on real cameras. Haven't seen any papers that use it for this.

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.

[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.