Analysis, Characterization, and Visualization of Freeway Traffic Data in Los Angeles

Scott Chacon
Analyst, Investment Banking and Capital Markets
Wells Fargo Securities
Email: scottchacon@gmail.com
Phone: 626-497-8986

Professor Alain L. Kornhauser
Director, Transportation Program
Department of Operations Research & Financial Engineering
Princeton University
Email: alaink@princeton.edu
Phone: 609-258-4657

Presented at the 53rd Transportation Research Forum Annual Conference
Tampa, Florida
March 14-17, 2012

Word Count: 6,137
Figures: 14
Tables: 3

March 10, 2012
Abstract

Presented is an analysis of a large volume of readily available loop detector based traffic data for the Los Angeles and Ventura Counties. The data suggests that the daily temporal variation of congestion along any directional road segment can be characterized quite well by a 10-parameter function. The function is shown to be suitable for use in the classification of road segments, such as having morning but no afternoon or evening congestion, as well as for the purpose of improving real-time forecasts of congestion ahead for use in generating dynamic real-time minimum estimated time-of-arrival turn-by-turn navigation instructions. Automation of the process allows for the characterization of all of the Los Angeles and Ventura Counties and can be applied to any metropolitan area having similar data. Several interactive and dynamic visualization tools using Google Earth are also developed and presented.
1 Introduction

Over the past two decades, more and more cities throughout the U.S. have recognized the value of real-time traffic monitoring to help understand and relieve congestion and make optimum use of a scarce roadway infrastructure. As a result, they have deployed a variety of sensors to capture the data. The data are used for many purposes, including its conversion into information delivered to the traveling public in many forms such as radio and television traffic reports, online traveler information systems, and the 511 Traveler Information System. Moreover, developers of in-car navigation systems are anxious to access this real-time traffic data to directly assist individual drivers on each and every trip. Ultimately they'd like to be able to generate turn-by-turn directions that continuously anticipate congestion ahead as the trip evolves. At this level, the data needs are challenging. What is required is a continuous real-time forecast of congestion for all possible routes. This is particularly daunting since it is well known that traffic congestion has three primary components. First, a recurring component due primarily to recurring, and thus predictable, variations in demand for travel in particular corridors. A second is associated with externalities such as weather and special events. Predictability of this element is tied to the predictability of the underlying causes. The third is due to low probability but highly disruptive shocks to the traffic system caused by incidents such as accidents, police actions, and emergency services. The inception of this third component is essentially unpredictable; however, the aftermath of an incident on traffic congestion is measurable as are the other two components through the use of real-time traffic monitoring systems.

1.1 Approaches to gathering traffic data

Traffic monitoring systems come in many forms. The most widely deployed is in the form of inductive loops that are embedded in the roadway at strategic locations. Chen et al. (2000) describes one such system for the state of California. These effectively measure the instantaneous speed of every vehicle passing over them. Thus they provide a thorough time history of both the speed and the volume of traffic at the location in the lane at which the loop is embedded. Similar forms of location specific speed monitoring include the use of radar and laser speed detection devices.

The primary shortcoming of these devices is that they provide speed only at a specific locations and no information about speed between those locations. This is acceptable only if one can reasonably assume how speed varies between the locations, say linearly. Ultimately, what is needed is the travel time on alternative paths throughout the roadway network. This is the integral of the inverse of the realized speed at each point.
along each alternative path. Without an appropriate assumption about how speed varies between detectors, one would need essentially an infinite number of detectors.

An alternative is to measure travel time directly between fixed points throughout the roadway network. This has the advantage of providing the direct measure of the primary performance of a roadway system; however, it requires the coordinated measurement of the passage of specific vehicles past multiple locations. Optical systems use machine vision to correlate the time of passage of specific license plates at deployed camera locations. Conceptually, these have the advantage of being able to monitor essentially any and “all” vehicles in the traffic stream; however, for technical reasons, not all license plates can readily be read in “all” environments. For example, they have severe difficulty at night. Other systems which tend to be free of environmental challenges but whose robustness is limited because they require some form of vehicle based device include: electronic toll tag, e.g. EZPass readers, cellular phone location record processors, bluetooth ID readers, and GPS “breadcrumb” data archivers. Toll tag readers are the most passive from a probe vehicle point of view but require the strategic and physical deployment of tag readers throughout the roadway system. The processing of cell sector exchange records normally archived by mobile phone service providers has the enormous advantage of placing essentially no demands on the probe vehicles other than the turning on of one’s cell phone and has an extremely large penetration rate in any traffic stream. Such systems can theoretically measure the time at which essentially all vehicles pass into or out of any cell sector. The disadvantages include the substantial difficulty in obtaining these data from the mobile telephone service providers and the correlation of cellular sector boundaries with the appropriate segments in the roadway network. GPS “breadcrumb” data processing holds enormous promise because of its data quality, precision and geographic robustness (it can monitor travel time on “any” roadway). It is severely limited because of its current low representation in any traffic stream. There exits no centralized mechanism for harvesting GPS sensor data. Most is used contemporaneously, then discarded. However, many in-vehicle navigation system developers have realized the value of the GPS breadcrumb data for customer support. These systems automatically archive what is actively harvested by the navigation companies. Those systems without cellular communications have simple easy to use batch harvesting procedures. Those operating on cellular phones archive these data to a central location on a real-time basis. This is becoming standard practice in the motor carrier industry where the data is communicated in real-time to centralized operations centers. Consequently, vast amounts of these data are becoming available at very low cost.
1.2 Research focus

This paper focuses on the characterization, visualization, and analysis of traffic congestion data. Focus is on the recurring aspects, in particular the time-of-day, day-of-week temporal aspects throughout the major arteries of major metropolitan areas. Theoretically, the data could come from any of the sources described above; however, the breadth and depth of the data requirements are such that today only loop detectors provide sufficient data for the proper characterization and analysis. Also, the application is centered on the Los Angeles and Ventura Counties because of the large number of loop detectors deployed throughout the basin as well as the availability of the data on the internet. The system, nicknamed PeMS, is a project conducted by the Department of Electrical Engineering and Computer Sciences at the University of California, at Berkeley.

The system receives data from loop detectors placed on most major freeways. Typically they are one-third to one-half a mile apart and may be used to measure traffic in a number of locations along a freeway. Detectors are situated on ramps, freeway-to-freeway connectors, high occupancy vehicle lanes (HOV), and in mainline lanes. Currently, PeMS maintains approximately 11,000 stations covering 5,400 miles of freeway across the Los Angeles and Ventura Counties. Figure 1 shows the location of all the loop detectors with their unique PeMS identification numbers displayed above their respective placemarks.

1.3 Loop Detector Data

Every thirty seconds the loop detectors send flow (the number of vehicles to pass over the loop during the time period) and occupancy (the proportion of the period that the inductance loop detected a vehicle above it) data to PeMS for processing and storage. The reported flow and occupancy data is aggregated over various time periods and then processed in order to compute speed. The data is aggregated into 30-second, 5-minute, 1-hour, or 1-day data files. Table 1 shows the general format of the raw (30-second) data files. A freeway “segment” is defined as the section of freeway between the midpoints between a series of three loop detectors. It is assumed that the traffic congestion characteristics are constant throughout each segment and characterized by the data captured (flow, occupancy, and implied speed) at the center loop detector.

<table>
<thead>
<tr>
<th>Table 1: Table Representation of 30-Second PeMS Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timestamp</td>
</tr>
<tr>
<td>MM/DD/YYYY, HH:MM:SS</td>
</tr>
</tbody>
</table>
One thing that affects the accuracy of this assumption is the distance between detectors - the longer the distance, the more opportunity there is for flow and occupancy to vary. Figure 2 shows the distribution of the distances between detectors for all “mainline” stations (excludes detectors located on on-ramps, off-ramps, freeway-to-freeway connectors, and HOV lanes). The distribution is skewed very heavily to the left, implying that most of the segments are less than one mile long. In fact, the average length of a segment is 0.70 miles, and the empirical standard deviation is 0.64; so, it’s fairly safe to assume that traffic is generally uniform over a segment.

Indeed, it would be nice to know the variation of speed within a thirty second period over one segment to further test this claim. Unfortunately, PeMS only reports thirty-second averages and not the individual readings that go into those averages. Figure 3 shows the percent change in speeds computed along one segment going from one 30-second reading to the next. The data is normally distributed with a mean very close to one. The tails are very thin, implying that speed changes between consecutive 30-second periods is generally very small, thereby further supporting the correctness of the assumption that flow and occupancy are constant over a freeway segment.
Figure 2: Histogram of Station Segment Lengths for Los Angeles and Ventura Counties

Figure 3: Histogram of Percent Change of Speed Between 30-Second Periods
2 Unit of Congestion Measurement

Anyone who’s ever driven on a Los Angeles freeway has probably complained about traffic. The system is nowhere near perfect, and commuters and travelers alike suffer because of the inefficiencies. But how does one quantify traffic? One way is to compare speeds to what is theoretically optimal. Jia et al. (2001) did exactly this. First, Jia et al. showed that the maximum throughput on Los Angeles freeways occurs at 60 miles per hour and concluded that anything below 60 mph (96 kph) is less than optimal and should be considered “congestion.” Delay is therefore defined as the additional vehicle-hours spent due to congestion, not being able to travel at a speed of 60 mph, with no marginal benefit of travel at higher speeds. The equation for delay is:

\[
\text{delay} = (\text{link length}) \cdot \text{flow} \cdot \max \left( \left[ \frac{1}{(\text{average speed})} - \frac{1}{(\text{target speed})} \right], 0 \right)
\] (1)

Delay has units vehicle-hours per time period, which it gets from the fact that link-length is in miles, flow is in vehicles per time period, and the inverse of speed is hours per mile. Figure 4 shows how delay is basically recurring but does include some day-to-day changes during a week. From Monday to Friday, recurring delays are large in the morning and smaller in the afternoon. Mondays and Thursdays seem to have the worst morning delays. The afternoon peak delays seem to decrease as the week progresses. Fridays have the least traffic overall. This suggests that there exists a recurring basic parametric time-of-day variation in congestion at each location that varies in intensity by day-of-the-week. On weekends, there is yet another pattern exhibiting very little delay - the morning peak is non-existent and midday and late-afternoon peaks are small for this road section.

2.1 Delay at Consecutive Stations

Figure 6 shows the delay at four consecutive mainline stations on the 10 freeway heading west. Their relative locations along the freeway are shown in Figure 5. The delay was computed using the 5-minute speed estimates from September 22, 2008. The first station that the traffic reaches as it travels west, station number 763330, is located about five miles west of downtown Los Angeles. It’s interesting to note that this station has significantly more delay compared to the other three. At almost every time, the delay decreases the further west the station is. This suggests that a bottleneck has formed upstream and dissipates just west of the first station. As implied in the beginning of the paragraph, the bottleneck’s formation is perhaps due to the focused inflow that occurs at station 763330 because of its closer proximity to more densely populated.
Figure 4: Separate Delay Time Series Plots for Five Consecutive Days

<table>
<thead>
<tr>
<th>Day of the Week</th>
<th>Actual Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Monday, August 22, 2008</td>
</tr>
<tr>
<td>2</td>
<td>Tuesday, August 23, 2008</td>
</tr>
<tr>
<td>3</td>
<td>Wednesday, August 24, 2008</td>
</tr>
<tr>
<td>4</td>
<td>Thursday, August 25, 2008</td>
</tr>
<tr>
<td>5</td>
<td>Friday, August 26, 2008</td>
</tr>
<tr>
<td>6</td>
<td>Saturday, August 27, 2008</td>
</tr>
<tr>
<td>7</td>
<td>Sunday, August 28, 2008</td>
</tr>
</tbody>
</table>
areas and freeway interchanges just west of downtown.

2.2 Animated Visualization of Delay

Using the capabilities of Google Earth, PeMS data was used to show delay trends over time. Colored rectangles were placed next to detector stations representing the current deviation of delay from the median. The edges of the rectangles were computed every five minutes for every station in the Los Angeles and Ventura Counties. Each square was given a time stamp, and the data was formatted into the Keyhole Markup Language (KML), which is the file format used by Google Earth. Rather than use the stations’ latitude and longitude as the corners of each of the boxes, points halfway between each station were used as the roadway anchors for the rectangles. Then the GPS locations of the two points at the top of the rectangle were computed using a bearing perpendicular to the roadway. For freeways heading north, the columns span east away from the freeway and for freeways heading south, the columns are to the west. Similarly, the columns stretch north for westbound freeways and south for those that are eastbound. The height of each rectangle was set at each time step such that the area would be proportional to the amount of delay at the particular station. Figure 7 shows two screenshots of an animation overlay for September 22, 2008. The figure is centered on the intersection of the 405 and 10 freeways in West Los Angeles. In the top subfigure, the polygons are very hard to see because the screenshot was taken at 3:03 AM when the traffic didn’t deviate from the median. This was included simply for the sake of comparison.

The bottom subfigure was taken later at 6:36 PM of the same day. The color of the green polygons signifies that delay at those stations was better than the median. Red or orange means the delay was worse than the median, and the longest columns are exhibiting delay that deviates the most (either positively or negatively) from the median. Like a frame-by-frame cartoon, the overlays if viewed in the order of their time stamps creates an animation showing how delay changes throughout the day. The evolution of bottlenecks
Figure 6: 24-hour Delay Time Series Plots for Consecutive Stations

<table>
<thead>
<tr>
<th>Station ID</th>
<th>Station Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>717029</td>
<td>1</td>
</tr>
<tr>
<td>717031</td>
<td>2</td>
</tr>
<tr>
<td>717033</td>
<td>3</td>
</tr>
<tr>
<td>763330</td>
<td>4</td>
</tr>
</tbody>
</table>

2008-09-22; First Station: 717029; Fwy: 10W/ML
Figure 7: Google Earth Delay-Polygon Overlay

(a) Early Morning:

(b) Rush Hour:
and their movement upstream on a freeway is very easy to see. Accidents also stand out because they look like bottlenecks with a downstream end that doesn’t move. Overall, the animations are very useful because the combination of color and spatiality gives viewers a better idea of what’s going on that can’t be garnered just from one or the other alone.

3 Predicting Weekday Delay

3.1 Method of Curve-Fitting

The most common delay graph typically has three humps: one representing the morning commute, one in the early afternoon, and one at the height of the afternoon commute. This shape suggests a graph (mathematical function) made up of a constant and three humps, say, normal distributions that is a function of ten parameters. Such a graph would be characterized by 10 parameters that would provide a best fit to the underlying loop detector data for any segment. As presented by Kornhauser et al. (2004), recurring time-of-day delay can be described by the following function:

\[
\text{delay at time } t = K + C_1 \cdot \varphi_{\mu_1,\sigma_1}(t) + C_2 \cdot \varphi_{\mu_2,\sigma_2}(t) + C_3 \cdot \varphi_{\mu_3,\sigma_3}(t)
\]

(2)

where \( \varphi(\mu, \sigma) \) is the pdf of the normal Gaussian distribution with mean \( \mu \) and standard deviation \( \sigma \). The explicit expression for the Gaussian pdf is:

\[
\varphi_{\mu,\sigma}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}
\]

(3)

In Equation (2), the values of \( \mu_i \) should be the locations of the heaviest congestion, with \( \sigma_i \) representing the lengths of the peak periods. The \( C_i \)'s are then scalars that translate the mean of the distribution vertically to match the value of delay at the peak times. To find the ten parameters, the curve-fitting exercise comes down to minimizing the sum of the squared errors between estimation function and the observed data points. The simple equation for the sum of the squared errors is:

\[
SSE = \sum_t (\text{delay}(t) - \hat{\text{delay}}(t))^2
\]

(4)

where \( \hat{\text{delay}}(t) \) is the observed delay at time \( t \). With the fitted curve, it’s a simple exercise to forecast delay forward because the expression is an explicit function for \( t \).
3.2 Delay Forecasting

Although forecasting delay based on a static model is an adequate approximation, a more precise approach is to update delay expectations as delay is observed throughout the day. That way people can know how severe congestion will get during rush hour while it’s still early in the peak period. Exponential smoothing is the preferred method for doing this. It takes the forecast estimate from the fitted curve and weights it against the live data. That way the most recent delay observations are incorporated into the model but not so much as to dominate. Weighting it against the estimation based on historical data serves to minimize the effects of abnormal peaks on the model. Kornhauser et al. (2004) presented exponential smoothing and forecasting equations for peak as well as non-peak times. The general equation for exponential smoothing during peak periods is of the form:

\[ S_n = \theta \cdot X_{n-1} + (1 - \theta) \cdot [\tau(t_n) - \tau(t_{n-1}) + S_{n-1}] \]  

(5)

where \( \theta \in \{0, 1\} \) is the smoothing factor and \( \tau(t_n) \) is the ten-parameter function fit to the data. At the times for which there isn’t yet data, meaning there is no \( X_{n-1} \), the \( \tau \) function is strictly used to predict the delay. The smoothing factor is simply equal to zero, and the forecasting equation becomes:

\[ S_n = \tau(t_n) - \tau(t_{n-1}) + S_{n-1} \]  

(6)

Equation (6) works just fine during peak periods, but it is slightly different for non-peak periods. There is an additional smoothing factor \( \gamma \in \{0, 1\} \) that is added to the model. The new smoothing equation becomes:

\[ S_n = \alpha \cdot X_{n-1} + (1 - \alpha) \cdot [\gamma \cdot S_{n-1} + (1 - \gamma) \cdot \tau(0)] \]  

(7)

The forecasting equation is also different because it has an added parameter \( c \):

\[ S_n = c \cdot S_{n-1} + (1 - c) \cdot \tau(0) \]  

(8)

The forecasting equation uses \( c \) to weight the previous smoothed value with the delay at time zero as predicted by the optimal curve, and the non-peak smoothing equation uses the new variable \( \gamma \) to weight the previously smoothed value and delay at time zero.
Figure 8: Station Number 717026 Average Workweek Travel Delay with Fitted Curve

Table 2: Ten-Parameter Function Coefficients for Aggregated Workweek Data at a Single Station

<table>
<thead>
<tr>
<th>K</th>
<th>$C_1$</th>
<th>$\mu_1$</th>
<th>$\sigma_1$</th>
<th>$C_2$</th>
<th>$\mu_2$</th>
<th>$\sigma_2$</th>
<th>$C_3$</th>
<th>$\mu_3$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0226</td>
<td>12.1755</td>
<td>8.3279</td>
<td>1.2446</td>
<td>2.9779</td>
<td>14.9270</td>
<td>1.7928</td>
<td>2.3202</td>
<td>18.6666</td>
<td>0.7872</td>
</tr>
</tbody>
</table>

(8:19 AM) (75 mins) (2:56 PM) (139 mins) (6:40 PM) (47 mins)

3.3 Application of Forecasting Method

A curve of the form in Equation (2) was used to fit the median delay across the seven days in Figure 4. The scatter plot of median values together with the optimal curve are shown in Figure 8. The curve follows the data very closely, showing a large spike in delay in the morning and a two-humped increase again in the afternoon. The curve parameters are listed in Table 2. The model predicts that morning rush hour occurs at roughly 8:19 AM, and the afternoon peak times are at 2:56 PM and 6:40 PM. The table gives values for the standard deviation of each normal pdf, which are measures of the length of the peak congestion periods.

From the table and the graph it’s clear that the morning peak is by far the most severe - it has $C_1$ that is four times that of $C_2$ and $C_3$. But the first afternoon peak period is by far the longest (139 minutes). Combined with the length of the later peak, which overlaps slightly, the afternoon congestion stretches over a much longer period. Because the second and third normals overlap, their interaction makes comparing the relative values of the first and the last two scalars misleading as to which peak(s) are higher. This could lead
to very odd results because the ten-parameter function is such that, for example, a wildly low value for $C_2$
could counteract a wildly high value for $C_3$, depending on the values of $\mu_2$ and $\mu_3$ and $\sigma_2$ and $\sigma_3$. Therefore
the results in tables like Table 2 must be taken somewhat lightly.

Figure 9 takes the fitted curve and uses data from another random day in September to forecast the delay
at three hour increments. The top two plots show that the data observations (green) during the morning
non-peak period followed the static fitted curve almost perfectly. In fact, it is difficult to tell which of the
lines is the forecasting curve and which is the static fitted curve. In the second two plots, the data points
began to stray away from the fitted curve during the beginning of the morning rush hour. The forecasting
curve (red) sits above the fitted curve (blue) throughout the rest of the day to show that the model predicted
that traffic would be worse than usual.

4 Determining Aggregate Delay Across all Los Angeles and Ven-
tura County Freeways

It’s difficult to determine just how bad the traffic congestion is on the Los Angeles and Ventura County
freeways and how this compares with other cities through the United States. Certain locations in the two
areas rank within the top ten worst corridors in the entire country according to a study done by Cambridge
Systematics for the American Highway Users Alliance in 2004. But could these few locations just be anomalies
and not the standard? This section uses the curve-fitting techniques discussed in Section 3.1 to characterize
the delay at all the stations in Los Angeles and compares the results across the seven days of the week and
holidays.

4.1 Overview of Analysis

The curve-fitting method described in Section 3.1 is expanded to include instances where a ten-parameter
function isn’t the proper way to model the traffic delay. There may be any number of daily peaks in traffic,
but this section will classify all stations as having at most three. The ten-parameter function discussed
earlier fits delay that has three peaks, but there may be two or one (or none in the case of constant delay).
Because of the massive amount of PeMS data, it was at first difficult to comprehend how to best go about
fitting the curves. The first step was to get a better understanding for the different types delay graphs that
existed across the entirety of the Los Angeles and Ventura Counties. A Google Earth visualization seemed
like the perfect solution to this problem. Utilizing the power of KML once again, time series plots were
Figure 9: Smoothed and Forecasted Delay using Three Hour Intervals
added to the information boxes that pop up when a placemark is selected. With a placemark for every station, by simply spanning across the counties and clicking on a placemark, one could explore the different types of delay plots by where they are located. Figure 10 shows the plot for one of the stations on the 10 freeway. For consistency, all of the plots are for the same day in September, and 60 mph was used as the target speed.

Exploring Google Earth showed that there were ten different types of delay graphs, each unique either because of the size, number, or location of peaks:

1. One AM: one peak before noon. Implies only a morning rush hour.

2. One PM: one peak in the evening. The reverse of the above; there is only an evening rush hour.

3. Large AM, Small PM: one bigger peak in the morning and one smaller peak in the evening. Morning and evening rush hour, with greater delay during the morning peak.

4. Small AM, Large PM: one smaller peak in the morning and one bigger peak in the evening. The reverse of the above; morning and evening rush hour, with greater delay during the evening peak.

5. AM & PM: one morning peak and one evening peak of the same size, distinguishable from the midday traffic. Commuters can expect the same amount of delay during morning and evening rush hours; midday delay is minimal.
6. AM, noon, & PM: one morning peak and one evening peak with heavy midday traffic. Same as the above except there is added midday delay.

7. All Day: bad traffic all day long. With the exception of 11 pm to about 4 am, these stations have traffic that is bad all the time.

8. AM & 2x PM: one morning peak and two in the afternoon. The type of delay fit best by the ten-parameter function described in Section 3.1; one morning rush hour and a longer, more spread out evening rush hour characterized by two peaks at different times in the afternoon.

9. 2x AM, PM: two morning peaks and one in the afternoon. The reverse of the above; morning rush hour has two peaks.

10. No traffic: no peaks that can be considered different from the baseline; what everyone hopes for.

One of each of the types of delay can be seen in Figure 11. These are simply characterizations of all the different shapes of delay graphs found. Later analysis will characterize each graph as one of the ten types and all stations exhibiting the same traits will be grouped into buckets where each bucket contains all the delay graphs of a single type.

4.2 Expanded Method of Curve-Fitting

The first step to fitting the curves is estimating the locations of possible peaks. This is done with the assistance of a built-in MATLAB function called “findpeak,” which given a vector will find the three largest local maxima. This is used as an estimate of the time location of the three humps in the $\mu_i$ in the ten-parameter function (Equation (2)). The difficult part is finding the $\sigma_i$ and $C_i$ because they affect each other. The maximum value of the normal pdf depends on the spread, and so the $C_i$ can’t simply be some scaling of the delay values of the local maxima. The solution begins with the following system of equations:

$$
\begin{align*}
y_1 &= K + C \cdot \varphi_{\mu,\sigma}(x_1) \\
y_2 &= K + C \cdot \varphi_{\mu,\sigma}(x_2)
\end{align*}
$$

where $y_1$ is the local maximum in the data and $x_1$ is the time at which this maximum occurs. The second point $(x_2, y_2)$ is used to give an estimate of the spread of the data around this peak. $y_2$ is an average of the 10 data points on either side of the maximum and $x_2$ is set to be ten data increments away from the peak.
Figure 11: Examples of Ten Different Types of Traffic Delay

(a) One AM:

(b) One PM:

(c) Large AM, Small PM:

(d) Small AM, Large PM:

(e) AM & PM:

(f) AM, noon, & PM:

(g) All Day:

(h) AM & 2x PM:

(i) 2x AM & PM:

(j) None:
If $\hat{D}(t)$ is the observed delay on one day where $t \in \{0, 0.0833, \ldots, 23.9167\}$, then $y_2$ is given by:

$$y_2 = \frac{1}{2T} \sum_{i=-10+0.0833}^{10+0.0833} \hat{D}(i)$$

and $x_2 = x_1 + 10 \times 0.0833$ (Note that five minutes is equal to 0.0833 hours and the data is in five-minute intervals; so, 0 corresponds to 0:00 AM (or midnight) and 23.9167 represents 12:55 PM). There is now a simple relationship between the $y_i$ and $x_i$, leaving only two variables that are unknown: $C$ and $\sigma$. There are also two equations in (9), making it easy to solve for the desired variables. The first step is to substitute each $\varphi_{\mu, \sigma}(x)$ with the expression given in Equation (3). Taking the natural logarithm of both sides in each gives:

$$\ln(y_i - K) = \ln\left(\frac{C}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}\right)$$

for $i = \{1, 2\}$. Multiplying the equation for $i = 2$ by $-1$ and using the law of logarithms to rearrange the variables yields the system:

$$\begin{cases} 
\ln(y_1 - K) = \ln(C) - \ln(\sigma \sqrt{2\pi} - \frac{(x_1 - \mu)^2}{2\sigma^2}) \\
-1 \cdot [\ln(y_2 - K) = \ln(C) - \ln(\sigma \sqrt{2\pi} - \frac{(x_2 - \mu)^2}{2\sigma^2})]
\end{cases}$$

Adding the equations in (10) gets rid of the second unknown, $C$, and the result reduces to:

$$\ln(y_1 - K) - \ln(y_2 - K) = -\frac{(x_1 - \mu)^2}{2\sigma^2} + \frac{(x_2 - \mu)^2}{2\sigma^2}$$

$$\ln\left(\frac{y_1 - K}{y_2 - K}\right) = -\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2}{2\sigma^2}$$

$$2\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2}{\ln\left(\frac{y_1 - K}{y_2 - K}\right)}$$

$$\sigma = \sqrt{\frac{-(x_1 - \mu)^2 + (x_2 - \mu)^2}{2 \cdot \ln\left(\frac{y_1 - K}{y_2 - K}\right)}}$$

Equation (11) is an explicit expression for one of the unknown variables, $\sigma$. Substituting for $x_1, x_2, y_1, y_2$ gives an estimate for $\sigma$ that can then be used in either of the equations in (9) to find the second of the unknown variables. If done for each peak found in a day of delay, this algorithm provides a very good
starting value for the parameters for a fitted curve of the form:

\[ D(t) = K + C_1 \cdot \varphi_{\mu_1, \sigma_1}(t) \cdot 1\{\text{There is at least one peak}\} + C_2 \cdot \varphi_{\mu_2, \sigma_2}(t) \cdot 1\{\text{There are at least two peaks}\} + C_3 \cdot \varphi_{\mu_3, \sigma_3}(t) \cdot 1\{\text{There are at least three peaks}\} \] (12)

where the values of the indicator functions are determined by the MATLAB “findpeak” function described in the beginning of the section.

4.3 Grouping Locations with Similar Daily Delay Characteristics

Section 4.2 introduced the basics of automating the process of fitting a curve to delay data from one station. The greater goal is to be able to easily fit curves for a large number of stations all at once. Using the methodology discussed in Section 4.2, this is done later for every mainline station in the Los Angeles and Ventura Counties, and subsequent analysis of the fitted curves allows one to broadly characterize the most prevalent types of delay across the entire area. The important questions being asked are: how many of each type of delay are there? Do most freeways only have heavy morning rush hours? Evening rush hours? How many areas are bad all day long? These are all answered herein with the help of the curve-fitting and new bucket assignment algorithm to be introduced now.

Automatically grouping similar fitted curves should be simple because of the form of the modified ten-parameter function used to fit the data (see Equation (12)). The assignment algorithm performs a different analysis depending on the number of parameters: one, four, seven, or ten.

1. When the curve-fitting algorithm returns a function of just a single parameter then delay is constant throughout the day and therefore should be put in the “All Day” or “None” bucket.

2. When there are four parameters (meaning the delay has at most one peak), the relative size of \( C \) scalar is compared to the constant \( K \) to determine whether the peak is a significant increase over the otherwise unchanging delay. If the \( C \) value is at least greater than two times the \( K \) value, then the delay is also designated as exhibiting no delay. A 2x multiple was found through trial and error to be the best at identifying peaks that should be recognized.

3. When there are seven or more parameters, the number of possible buckets increases dramatically and so must the analysis. The first step is to identify the locations of the peaks. Since there is no bucket
for instances when there are two peaks either before or after noon and none in the other half of the day, this is the bucket that the algorithm searches for first because it is the easiest: if the $\mu$ values are on the same side of noon, then the peaks are treated as one and the curve is assigned to either bucket one (one peak before noon) or two (one peak in the evening). If this first check doesn’t assign the curve to a bucket, meaning the peaks are on both sides of noon, then the curve could either have two peaks on either side of noon, one peak in the morning, or one peak in the evening. If the first derivative (computed using the Euler Method) of the delay function has zeros on either side of noon, then the “One AM” and “One PM” buckets are ruled out unless one of the peaks is less than 2x the $K$ value for the constant delay.

4. When there are ten parameters, the realm of possibilities increases because any number of the detected peaks could in fact not hold up under greater scrutiny. In the bucket assignment algorithm, the first check is to check the legitimacy of all of the peaks by comparing the local maxima to the baseline. Second, the locations of the peaks relative to noon are established. If all of the peaks are located on one side of noon, then once again all the peaks are treated as one and the bucket is either “One AM” or “One PM.” If the function detects a negative $C$ scalar, then the derivative of the curve is used to determine the number of positive peaks and the analysis continues as if there were seven parameters to start off with. If a curve hasn’t been assigned to a bucket after the analysis discussed immediately above, then the curve has proved to have either two or three peaks. At this point the functions could have either seven or ten parameters. For curves with ten parameters, the only way there aren’t three peaks is if one set of parameters doesn’t add anything substantial to the model, and this is test is the next to be done. A “what-if” analysis is done to compare the total delay (area under the entire curve) with and without the different sets of parameters. If the difference between the delay is less than four vehicle-hours (also found through trial and error), then the number of effective peaks is reduced from three to two.

Table 3 shows the parameters for one station on the 10 freeway west. From left to right, the columns represent Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, and a holiday. It should be clear now from the results that the parameters can’t tell the story themselves. Clearly a good bucket assignment algorithm is important because the curve-fitting function can return unusual results. In the table, some of the $C$ scalars are negative, some of the $\mu$ values are the same, and some of the peaks aren’t much bigger than the constant $K$. When looking at the raw coefficients, the important thing to remember is that the normal distributions in the 10-parameter function (Equation (2)) may overlap depending on their
means and variances. By interacting with each other, one normal factor in the function might counteract or augment the effects of another; so, coefficients will only make sense when taken together. Take the Saturday coefficients as an example: the values of the second and third normal factor are almost identical - $\mu_1 \approx \mu_2$, $\sigma_1 \approx \sigma_2$, and $C_1 \approx -C_2$. Because $C_1$ is the opposite of $C_2$ and the two normals are centered and spread nearly identically, the effect of one on the overall shape of the curve completely negates the effect of the other. Some sets of coefficients might on the surface imply completely different curve shapes; however, the raw numbers can be misleading because of the nature of functions with multiple overlapping normal distributions. As another example, take the coefficients for Wednesday and Thursday: Wednesday has two sets of parameters while Thursday has three. From the $\mu$ values for Thursday, it would appear that there are three peaks because of their separation. However, Figure 12 shows that these sets of parameters lead to almost identical plots.

In Figure 12, there is a clear difference between mid-week, weekend, and holiday travel. From Monday to Friday, there are two peak periods, one in the morning and the evening, and the morning rush hour is the worst. On Tuesday and Friday there is one morning peak and two peaks in the afternoon, but the traffic isn’t the same just because of this commonality. The smaller spread of the morning peak on Friday, a measurement of length of the morning rush hour, shows that fewer people go into work, and the increased separation of the two afternoon peaks shows that people don’t stay as late either. The standard deviation of the morning peak on Friday is 7.49 (see Table 3), which is less than the 8.05, 8.54, 8.35, and 8.41 standard deviations of the morning peaks on the other weekdays.

It’s clear from the last row of Table 3 that the algorithm provides fitted curves with varying qualities.

Table 3: Mid-week, Weekend, and Holiday Fitted-Curve Parameters for Station 717031 on the 10 Freeway Heading West

<table>
<thead>
<tr>
<th>Date</th>
<th>9/22/08</th>
<th>9/23/08</th>
<th>9/24/08</th>
<th>9/25/08</th>
<th>9/26/08</th>
<th>9/27/08</th>
<th>9/28/08</th>
<th>7/4/08</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>-42.16</td>
<td>0.18</td>
<td>0.23</td>
<td>-0.08</td>
<td>0.28</td>
<td>-0.38</td>
<td>0.02</td>
<td>-0.57</td>
</tr>
<tr>
<td>$C_1$</td>
<td>47.82</td>
<td>31.20</td>
<td>28.55</td>
<td>31.70</td>
<td>13.32</td>
<td>13.09</td>
<td>0.17</td>
<td>2747.58</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>8.05</td>
<td>8.54</td>
<td>8.35</td>
<td>8.41</td>
<td>7.49</td>
<td>14.16</td>
<td>2.00</td>
<td>8.01</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>2.07</td>
<td>1.47</td>
<td>1.38</td>
<td>1.51</td>
<td>0.69</td>
<td>1.37</td>
<td>0.84</td>
<td>1.59</td>
</tr>
<tr>
<td>$C_2$</td>
<td>12843.90</td>
<td>7.77</td>
<td>9.93</td>
<td>7.98</td>
<td>11.63</td>
<td>10906.64</td>
<td>3.41</td>
<td>-2740.46</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>12.53</td>
<td>17.74</td>
<td>18.24</td>
<td>15.28</td>
<td>14.84</td>
<td>14.16</td>
<td>13.83</td>
<td>8.01</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>124.06</td>
<td>1.21</td>
<td>1.07</td>
<td>4.17</td>
<td>1.49</td>
<td>4.15</td>
<td>0.68</td>
<td>1.59</td>
</tr>
<tr>
<td>$C_3$</td>
<td>12.38</td>
<td>3.51</td>
<td>9.23</td>
<td>6.06</td>
<td>-10894.09</td>
<td>3.62</td>
<td>23.75</td>
<td></td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>17.97</td>
<td>18.95</td>
<td>18.47</td>
<td>18.41</td>
<td>14.16</td>
<td>15.91</td>
<td>10.12</td>
<td></td>
</tr>
<tr>
<td>$\sigma_3$</td>
<td>3.1</td>
<td>0.35</td>
<td>0.93</td>
<td>0.66</td>
<td>4.14</td>
<td>0.57</td>
<td>9.19</td>
<td></td>
</tr>
<tr>
<td>SSE</td>
<td>589.20</td>
<td>64.13</td>
<td>86.24</td>
<td>64.40</td>
<td>95.93</td>
<td>54.40</td>
<td>59.60</td>
<td>30.40</td>
</tr>
</tbody>
</table>
Figure 12: Eight Days of Fitted Curves for Station 717031
of fit. The Monday curve, for example, doesn’t appear to fit the data well based on the SSE value. For the rest of the days of the week and the holiday, the algorithm does much better. On a data set with 288 points (every 5-minutes for 24 hours), a sum of squares of 70.00 represents on average 0.24 vehicle-hours per 5-minute period difference between the estimated and observed delay. Depending on the location, delay could vary between 20 and 0 over a single day; so, the conclusion is that the fitting algorithm does a good job of capturing the movements of the data.

4.4 Prevalence of Different Types of Delay

In order to get an idea of how different types of traffic zones are distributed around the counties, curves were fit to delay data for Monday, September 22 from each of the 1,500 mainline stations, and afterwards the curves were assigned to buckets. Figure 13 shows the locations of eight out of the ten types of traffic classifications. “One AM” delay and “One PM” delay are pretty evenly distributed throughout the city. The sets seem to overlap, which is most likely a result of commuters all traveling one direction on the freeway in the morning and then returning on the same freeways in the evening. “Large AM, Small PM” delay (one bigger peak in the morning and one smaller peak in the evening) doesn’t seem to be very widespread - there are very few stations exhibiting this form of delay, and they seem to be mostly around the periphery. In contrast, “Small AM, Large PM” delay (one smaller peak in the morning and one larger peak in the evening) is located all throughout downtown and the surrounding area. This is consistent with bi-directional commuting, meaning that some people live where others work while the other group works where the others live. Logically this would occur most often in the more densely populated areas, which are mostly in and around downtown. The number of people commuting one way in the morning isn’t the same as the number traveling in the opposite direction because the sizes of the peaks differ. It’s interesting to note that there are in fact very few locations where the morning and evening peaks are the same. Similar to “AM & PM” delay (equal peaks) discussed, Type 6 delay has equal morning and evening peaks but with added midday traffic - these are spread throughout the city. Stations exhibiting this type of delay are on stretches that have heavy rush hours and midday travel caused either by trucks or other business travelers. The last two delay characterizations, “AM & 2x PM” and “2x AM & PM” (two peaks on one side of noon and one on the other), which are opposites of each other, cover Los Angeles in no special way. It’s not surprising that “AM & 2x PM” (one morning peak and two in the afternoon) is spread all over the city because it’s supposed to be the most common. The interesting thing is that there seem to be just as many of “2x AM & PM,” which has two morning peaks and one in the evening. Two possible explanations for the two morning peaks is that
some people leave earlier than others in order to avoid the heavier traffic, or they are required to report to work early.

4.5 Bucket Composition Changes Across Days of the Week

The ten-parameter function from Equation 2 was used initially to model delay because it was believed that traffic had three peak periods. This section has shown that in reality most delay graphs don’t have three peaks. Figure 14 shows the percentage of each type of delay on each of the eight days of the week. With the exception of Monday, there is no noticeable difference in the workweek histograms. The first, second, and fourth buckets are most common, followed by the third and then the eighth and ninth. In addition, there are very few delay graphs of either “AM & PM,” “AM, on, & PM,” “All Day,” or “None” in comparison.

On Monday, the ninth bucket is by far the biggest, containing close to 80% of the delay graphs. Starting on Saturday, the distribution changes completely. The number of stations exhibiting “One PM” delay increases greatly, perhaps representing a tendency for people to only travel in the evening. This goes for Sunday as well except that on Sunday this is almost the only type of delay. The holiday histogram looks a lot like that for one of the Tuesday through Friday weekday plots. This shows that people still do a lot of traveling even if they’re off work. It makes sense because like Christmas Eve, the 4th of July is a time for people to throw parties, creating transportation demand. The interesting thing is that the get-togethers seem to attract people generally in the same timeframe as a job. One would’ve expected a lot more evening round-trip travel because that’s the most typical time for a party. The histogram shows, however, that people travel constantly throughout the day on the 4th just as they would during the week.

5 Conclusion

This paper explored various methods of visualizing and characterizing freeway traffic data. Through the analysis of traffic trends and the characterization of traffic delay on a large scale, this paper provides insights to the severity and breadth of traffic in the Los Angeles and Ventura Counties. Empirical delay curves were fit using complex functions and the least squares method. The resulting functional conclusions were assigned to one of ten different categories, each representing delay curves with the same general attributes. It was shown that while conventional theory proposes that delay graphs typically have three peak periods, delay curves with one peak (either before or after noon) are more common in the area studied.
Figure 13: Visualization of Delay Curve Classifications for Every Mainline Station in the Los Angeles Basin

(a) One AM:
(b) One PM:
(c) Large AM, Small PM:
(d) Small AM, Large PM:
(e) AM & PM:
(f) AM, noon, & PM:
(g) AM & 2x PM:
(h) 2x AM & PM:
Figure 14: Histograms of County-wide Delay Classifications by Day of the Week
References


