CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Freeway Performance Measurement System (PeMS)

Chao Chen

California PATH Research Report
UCB-ITS-PRR-2003-22

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation; and the United States Department of Transportation, Federal Highway Administration.

Report for Task Order 4301

July 2003

ISSN 1055-1425

Freeway Performance Measurement System (PeMS)

by

Chao Chen

B.A. (University of Virginia) 1994
M.S. (University of California, Berkeley) 1996

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Pravin Varaiya, Chair
Peter Bickel
Alistair Sinclair

Fall 2002

The dissertation of Chao Chen is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2002

Freeway Performance Measurement System (PeMS)

Chao Chen

Abstract

Freeway Performance Measurement System (PeMS)

by

Chao Chen

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Pravin Varaiya, Chair

The freeway Performance Measurement System (PeMS) collects real time traffic data from sensors and generates performance measures of vehicle miles traveled, hours traveled, and travel time. This project is sponsored by the California Department of Transportation (Caltrans). PeMS provides tools and reports for traffic planners, operators, and engineers. It has a Web interface.

Growing traffic demand in metropolitan areas has far outpaced increases in freeway lane-miles in the United States. The solution to congestion lies in increasing the efficiency of existing infrastructure. Performance measurement is the first step in effective management and operation of any system. Currently, the freeway system is not managed scientifically. Planning and operating decisions are made without accurate knowledge of the performance of each part of the system. PeMS collects data from automatic sensors that are already installed on most of California freeways. Its large database of real time and historical data

allows us to accurately measure the performance of freeways and its trends. Traffic planners need this information to allocate the available resources to improve mobility.

PeMS computes performance measures and other traffic quantities from sensor data. Among them are speed, vehicle-hours of delay, vehicle-miles traveled, and travel time statistics. These values can be visualized in plots and summarized in reports, and they are available online through a Web interface. Policy makers can use PeMS to evaluate the effect of their decisions and set performance targets, planners monitor trends in congestion and respond with congestion-reduction measures, engineers view detailed data to improve conditions at specific locations, and travelers use the information to make more informed decisions.

Researchers use PeMS's database to analyze traffic behavior on a large scale. We present some results from studies on freeway capacity, travel time variability, and the impact of incident on overall delay. In these cases, using observations from a large number of locations and times allows us to characterize traffic flow statistically.

PeMS processes raw data into useful forms. It computes speed from single loop detectors, predict travel time from real time and historical data, and detect and fix data errors. We describe these data processing algorithms, which are based on empirical models and fitted to historical data.

Pravin Varaiya
Dissertation Committee Chair

To my wife

Contents

List of Figures

List of Tables

Acknowledgments

I thank my advisor Pravin Varaiya for his guidance over the past five years and on this thesis. He had many students and was busy even on weekends, but he was always available to discuss a problem and offer insights and suggestions. He set high expectations and allowed me to reach my potential. Thanks to Karl Petty for teaching me anything I wanted to know about computers and technology. Thanks to Professor Alex Skabardonis for his transportation expertise and advice.

Erik van Zwet contributed significantly to this thesis in the areas of travel time prediction and speed estimation. Jaimyoung Kwon made important contributions on error detection and imputation. Thanks also to Professors Peter Bickel and John Rice for their insightful suggestions on algorithms, and to Professor Alistair Sinclair for his help on shortest path routing.

PeMS is supported by grants from Caltrans to the California PATH Program. I am very grateful to engineers from Caltrans Districts 3, 4, 7, 8, 11 and 12 and Headquarters for their encouragement, understanding, and patience. They continue to shape the evolution of the PeMS vision, to monitor its progress, to champion PeMS within Caltrans. Without their generous support this project would not have reached such a mature stage. In particular, I thank Hamed Benouar, John Wolf, and Fred Dial of Headquarters for their support of PeMS, Laurie Guiness and Ali Grabel for their help with GIS data, Martha Styer and Ahoura Vahedi of Caltrans and Julian Camacho, Tommy Iem and other employees of Camco Communications for their assistance on loop detector maintenance.

Thanks to Jennifer Moriarta for her support and editing.

The contents of this paper reflect the views of the author who is 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.

Part I

Applications And Specifications

Chapter 1

Introduction

The Freeway Performance Measurement System (PeMS) is an intelligent transportation management tool. It collects real time traffic data from sensors, stores them, and makes them available on-line. From its rich database of historical data, PeMS generates standard, well-defined performance measures of the freeway system. Trends in congestion can be tracked, locations of growing gridlock can be detected, and the service quality provided to drivers can be monitored. Another important use of these data is in providing real time route guidance and travel time predictions.

State Departments of Transportation (DOTs) operate the freeway system. Increasingly, transportation planners are coming to the realization that highway construction alone cannot solve the nation's congestion problem. Figure 1.1 shows the trend in delay, vehicle-miles traveled (VMT), and lane-miles of freeway in the Los Angeles region between 1982 and 1990 [2]. While the number of lane-miles increased a modest 30%, vehicle-miles traveled increased by 70% and delay increased by 270%. This trend is also observed nation-

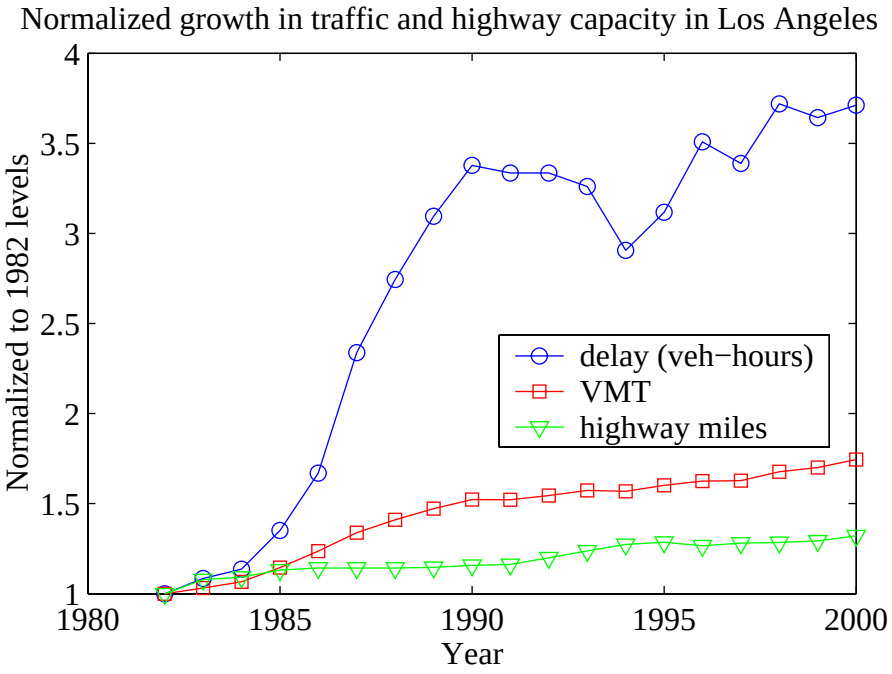


Figure 1.1: Congestion delay, vehicle-miles traveled (VMT), and lane-miles of freeway in Los Angeles, normalized to 1982 levels.

wide. The growth in total vehicle-hours of delay on in metropolitan areas far outpaced the growth in lane-miles of freeway. Between 1982 and 2000, the average annual delay per peak road traveler grew from 16 hours to 62 hours, a 288% increase [2]. During the same period, the total number of miles of freeway grew by only 35% [3].

From the above statistics, it's clear that congestion cannot be alleviated by increasing capacity alone. In many metropolitan areas, there is simply no more room to expand the freeway system. The solution must come from a combination of smart development, increased transit and car-pooling, and increasing the efficiency of existing infrastructure through information technology and intelligent management [2]. The last of these options is the most promising. Recent advances in telecommunications, electronics, computing, and the Internet have made it possible to build an intelligent transportation management

system that monitors traffic and provides performance measures.

To efficiently provide transportation service to the public, the DOTs need to know the freeway system's performance. Several important performance measures are total vehicle hours traveled (VHT) and vehicle miles traveled (VMT), and travel time variability. A good system needs to provide high VMT to VHT ratio and provide a high level of service. Access to the current and historical values of these measures allows us to evaluate management strategies and optimize freeway performance.

Performance measurement is a priority of Departments Of Transportation. The Highway Congestion Monitoring Program (HICOMP) of the California DOT (Caltrans) produces an annual report that documents congestion levels in each of its 12 districts. This report publishes the total delay in vehicle-hours by location in each district. It also gauges the service quality by measuring travel times on key routes, and the percentage of freeways in congestion. Usage is measured in the report of Annual Average Daily Traffic (AADT), which documents the annual flow rate at locations along freeways.

The measures above do not directly measure the performance of the freeway system in producing "output" and consuming "input". In contrast, PeMS defines several natural, meaningful, and computable performance measures. It measures the output of the freeway system in total vehicle-miles traveled, and input in vehicle-hours traveled. PeMS also computes the cost of congestion in delay, a consistent measure of average speed, as well as travel time and its variability. Current data collection methods are manual, labor intensive, and sporadic. Results based on these data are inaccurate. PeMS results much more accurate and detailed because it uses a large amount of detector data from thousands of locations

covering most California metropolitan areas. Users access PeMS online using Web-based applications, and generate performance measures and reports of current and historical conditions on-command. In contrast, existing congestion reports are generated annually. The timeliness of PeMS's results allows Caltrans to monitor short term and long term trends and respond quickly to changing conditions. PeMS's vast historical database and algorithms enable users to perform statistical analysis in order to understand the stochastic nature of traffic flow.

In addition to performance measurement, PeMS provides a suite of on-line applications that help traffic engineers diagnose specific congestion problems. PeMS provides the ability to visualize data at many temporal and spatial scales. For example, PeMS applications can be used to plot the daily delay on a given to show trends and outlier days which may be of interest. Data from a selected day can be used to analyze a congestion event by plots in time and space or contour plots in two dimensions. Real time data displays and automated alerts allow the freeway system operator to respond to accidents and advise drivers of congestion.

Recent developments in computer and information technology make PeMS possible and practicable. PeMS is designed to be accessed on the Internet through a Web browser, so its use does not depend on the availability of proprietary local clients. It's developed with an open software architecture, which makes it easy to add more capabilities as needs arise and technologies improve. Traffic data are collected from Caltrans loop detectors via district Traffic Management Centers (TMCs). These detectors measure average flow rate and freeway occupancy at thousands of locations in California. We also collect incident

information from the California Highway Patrol (CHP) control centers. We designed the system to collect data from as many sources as possible, and organize the different types of data so they can be correlated with one another and studied together. All data are stored in an Oracle database on-line, which currently has two terabytes and grows by two gigabytes per day.

PeMS does far more than simply store and disseminate data. It is impossible to realize the huge amount of information available by visually examining the raw data. PeMS's many applications retrieve the relevant data and present them in various levels of aggregation and in many types of views. Furthermore, many important traffic quantities are not directly measured and must be calculated. The algorithms that perform these calculations are based on statistical models. For example, we observed a linear dependence of future travel times on current travel time, which became the basis for our travel time prediction algorithm. From the measured occupancy and flow, we are able to compute speed, vehicle-hours traveled, vehicle-miles traveled, travel time, and more.

Performance measurement benefits travelers by allowing them to make informed decisions, such as with the use of Automated Traveler Information Systems (ATIS). Using real time data, such a system alerts drivers of incidents, indicates areas of congestion, calculates alternate routes, and predicts travel times. These services are natural by-products of the data and processing already a part of PeMS. We demonstrate a travel time prediction and routing Web-based service for Los Angeles.

PeMS's results are derived from measurements based on empirical models. The rich database of PeMS provides many opportunities to make statistical characterizations of

traffic phenomena. With this vast amount of data also comes the challenge of data quality assurance. Because of the large number of detectors and the harsh environment in which they operate, there are many malfunctions, introducing missing and invalid samples. A significant part of PeMS is devoted to diagnosing detector health and imputing data. On the one hand, PeMS detects bad sensors and generates detector health reports so they can be repaired. On the other hand, PeMS replaces bad samples with imputed values from neighboring good sensors using a probability model.

The body of this thesis begins with a description of PeMS Web-based applications in Chapter 2. Here, we define PeMS performance measures and explain their roles in transportation management. It is followed by descriptions of PeMS applications, which are explained in a few examples. An ATIS application and loop diagnostics applications are also described. Chapter 3 demonstrates the potential of PeMS for academic research. It highlights three statistical studies on capacity analysis of bottlenecks, travel time variation in Los Angeles, and the impact of incidents on overall delay. These studies characterize some important features of traffic using data from many locations and times. In Chapter 4, we describe the components that make up PeMS, including data collection, sensor technology, database structure, and the system architecture of PeMS applications. Chapter 5-7 contain several PeMS algorithms that correct data errors, calculate speed from single loops, predict travel times, and find shortest routes.

People and their contributions

PeMS is a collaborative effort involving many researchers. It was conceived by Professors Pravin Varaiya and Alex Skabardonis, with the support of Caltrans Headquarters. Since then, researchers from several different departments at UC Berkeley have contributed to its development. Because this thesis is a comprehensive description of the entire system, it includes the work not only of the author, but also of all the other members of PeMS. The following is a list of main contributors to PeMS.

Area	People
Software and hardware architecture, and data networking	Karl Petty, Pravin Varaiya
Performance measurement	Pravin Varaiya, Alex Skabardonis
Data quality	Peter Bickel, John Rice, Jaimyoung Kwon, Erik van Zwet, Zhanfeng Jia, Chao Chen
Single-loop speed algorithm	Ben Coifman, Zhanfeng Jia, Chao Chen, Erik van Zwet
Travel time prediction	Xiaoyan Zhang, Erik van Zwet, Chao Chen
Routing	Alistair Sinclair, Chao Chen

Table 1.1: Contributors to PeMS research.

Chapter 2

PeMS Applications

2.1 Performance measures

PeMS provides performance reports and congestion analysis tools for DOT planners, operators, and engineers. To appreciate these tools, it's helpful to understand what the DOT does. The California Department of Transportation (Caltrans) designs, builds, maintains, and operates the state's highways [4]. The highway transportation system has an important impact on the nation's economic strength and efficiency [5]. Although its precise contribution to the economy is hard to measure, the highway system facilitates the movement of raw materials and manufactured goods, and enables people to get to work and conduct business. Therefore, the DOT devotes significant resources to maintain its efficient operation. To manage the 15,000 miles of highways in California, Caltrans employs 23,000 people, and had a budget of $6.8 billion in 2000-2001. Most of the budget ($4 billion) went to capital outlay projects, such as repaving roads[6].

The goal of the DOT is to provide the efficient movement of goods and people.

To achieve its goal, the DOT needs to measure its performance. By monitoring system performance over time, the DOT can determine whether its goals are met and address the weaknesses. Performance measures provide accountability for investments in transportation programs and point to priorities in funding projects. They also provide information for drivers to make informed decisions about route choices and travel behavior [7].

Although the transportation system has many impacts on the economic, health, and safety well-being of the public, it is not practical to try to maximize these impacts from an operational point of view [8]. For operations, we need to use measurable quantities that are directly impacted by management decisions. The direct "output" of the freeway system is the movement of goods and people. Therefore, total number of vehicle-miles traveled (VMT) is a well-defined quantity that measures the system output. The system input includes the fixed cost of the DOT's budget, and also the variable input of people's time, fuel costs, and environmental impacts. Of these, the most significant and measurable is the time spent traveling. The total travel time is measured by vehicle-hours traveled (VHT). An efficient transportation must have a high ratio of VMT to VHT.

2.1.1 Current performance measurement

Caltrans produces the Highway Congestion Monitoring Program (HICOMP) report [9] and the Annual Average Daily Traffic (AADT) report [10] each year to measure freeway system performance. The HICOMP report measures delay by county and district; it also reports the most congested locations. For a trip, Caltrans defines delay as the difference between actual travel time and the travel time of the same trip at a constant speed

of 35 mph, or zero if the difference is negative,

$$\text{delay}_{Caltrans} \stackrel{\text{def}}{=} \max\left\{\text{travel time} - \frac{\text{length of trip (in miles)}}{35 \text{ mph}},\ 0\right\}.$$

For example, the 1998 HICOMP report for the San Francisco Bay Area showed 112,000 vehicle-hours of average daily delay, versus 60,000 vehicle-hours for 1994. The AADT report, on the other hand, measures the usage of the freeways. It reports the average daily volumes at locations along the freeway. For example, on I-710 in Los Angeles at the junction of I-105 (postmile 18.44), the average daily traffic volume was 206,000 vehicles per day (in both directions) in 1999, and 188,000 in 1992, an increase of 9.6%. Using these measures, Caltrans planners and operators decide how to allocate resources to improve performance.

There are several problems with the HICOMP and AADT reports. In the case of HICOMP, the delay measurements are based on a reference speed of 35 mph. The delay would be different if another reference speed was used. It would be better to also measure the VHT, because this is the absolute amount of time spent and doesn't depend on a reference speed. In the case of the AADT report, total flow rates are measured. But flow rates from different locations cannot be easily combined to give regional statistics – it doesn't make sense to talk about the "average flow rate" of Los Angeles. By comparison, VMT is a consistent measure because the VMT of different segments are additive to provide meaningful aggregates, as we will see later in this chapter. Another shortfall of the HICOMP and AADT reports is that they only give annual figures, which are actually based on very few data points collected over several days or weeks only. Therefore the results cannot capture the variability of freeway conditions and are likely to be misleading. Also, annual reporting cannot detect changes on a smaller scale.

2.1.2 PeMS performance measures

PeMS defines several basic measures of performance listed in Table 2.1. These measures are well-defined, interpretable, meaningful to transportation system performance, and computable from available data. VMT is a meaningful measure of the system's output.

Measure	Description	Units
VMT	vehicle-miles traveled	vehicle-miles
VHT	vehicle-hours traveled	vehicle-miles
T	travel time	minutes
D	delay	vehicle-hours
S	average speed	miles per hour

Table 2.1: PeMS performance measures.

If 1 million people used the freeway system to travel 10,000 miles each in one year, then the total VMT of the system in the year is 10 billion vehicle-miles. If all of the miles were traveled at 60 mile per hour, then the annual VHT would be $1/6 = 0.65$ billion vehicle-hours. This is the cost (in time) of traveling those miles. These are the direct output and input of the system, therefore they are the most useful quantities to know for effective system management.

VMT and VHT are defined for any time period, and for any collection of segments, such as a freeway or an entire metropolitan region. If a freeway has segments $1, 2, \ldots, n$, and $\text{VMT}_i(t)$ is the VMT of the ith segment during time t, then the VMT of the freeway is

$$\text{VMT}_{freeway}(t) \quad = \quad \sum_{i=1}^{n} \text{VMT}_i(t).$$

VHT of the freeway is also the sum of the segment VHT's. The additive property of VMT

and VHT means we can compute hourly, daily, monthly, and annual measures of their performance and observe their trends.

Travel time is another important performance measure. While VMT and VHT capture the aggregate performance of the system, travel time and its variability measure the service quality received by the users. Travel time is defined for any trip, which includes a route and a departure time. For a given time period, the total travel time of all trips equals the total VHT. However, the uncertainty in travel time is an extra cost that VHT does not measure. Suppose the average travel time on a certain route is 25 minutes, but individual trip travel times vary randomly and can be up to 40 minutes. Then, to guarantee on-time arrival, a driver has to budget 40 minutes for each trip. An extra 40-25=15 minutes is lost on each trip because of the unpredictability of travel time. We present detailed travel time statistics in Chapter 3 and their computation.

The average speed of a single trip is a well-defined measure of its "efficiency." Similarly, the average speed of all trips of the system during t is also well-defined by

$$S(t) \stackrel{\text{def}}{=} \frac{\text{VMT}(t)}{\text{VHT}(t)}. \tag{2.1}$$

PeMS uses S to measures the efficiency of a freeway system at generating VMT.

PeMS computes a fifth performance measure D for delay. D is defined similarly to the Caltrans definition,

$$D \stackrel{\text{def}}{=} \max\left\{\text{VHT} - \frac{\text{VMT}}{v_r},\, 0\right\}, \tag{2.2}$$

where v_r is a reference speed that can be specified by the user. Caltrans uses $v_r = 35$mph. Delay is useful because it separates the total VHT into the minimum travel time and extra

travel time above the minimum. PeMS proposes using $v_r = 60$mph, the free flow speed. Using this v_r, delay is the cost of congestion in extra travel time.

The performance measures above are natural and meaningful gauges of system performance. A good transportation system should produce high VMT to low VHT, with predictable travel times, low delay, and high average speed. PeMS computes performance measures from sensor measurements of volume, occupancy, and speed automatically collected from many locations on the freeway. The following section discusses the need for automatic data collection.

2.1.3 Data collection methods

Currently, there are several methods of collecting traffic data. Caltrans uses the floating car method to collect data for the HICOMP report. Professional drivers repeat trips on selected routes at 15-30 minute intervals while equipment in the probe vehicles record the travel times of each trip. A route typically consists of several freeway segments several miles long. The probe runs are made only during peak hours and only on Tuesday through Thursday of the week to eliminate weekend effects. If a probe vehicle trip encounters an incident, data from that trip is discarded. Therefore only data about recurrent congestion are collected. The travel time measurements are used to estimate the total delay in vehicle-hours. This is found by multiplying the average delay of the probe vehicles by the estimated traffic volumes at the test locations [9]. Probe vehicle runs are labor intensive and are only performed twice a year, each time for several days for each route. Annual measures are extrapolated from these results. Data used to produce the AADT report are collected using mobile electronic equipment that is moved from location to location [10]. Therefore, only

a small number of samples are collected from each location. Annual figures are once again extrapolated based only on a small number of measurements.

While existing data collection methods are often labor-intensive and sporadic, PeMS collects data automatically from 23,138 loop detectors, covering all of the state's major metropolitan areas. These detectors measure the vehicle count and occupancy for each 30-second period. While some detectors are double loops, which also measure average speed, most are single loops. PeMS uses an algorithm to estimate speed from single loop measurements described in Chapter 6. We give a description of loop detectors and the data collection process in Chapter 4. We will show that traffic conditions at a location are highly variable on different days, so measurements made on only a few days cannot capture the real behavior. Similarly, congestion varies widely by location. Only a comprehensive coverage of detectors can accurately measure regional conditions. Furthermore, real-time monitoring can be achieved only with automatic data collection. Loop data are sent over phone lines to a central computer and then to PeMS on Caltrans's Wide Area Network (WAN).

There are many other types of traffic detectors, such as ultrasonic, microwave, and laser sensors, processed video, magnetometers, and tube-type vehicle counters. These sensors are mounted at fixed location and measure different combinations of speed, vehicle count, and occupancy [11]. Travel time can be measured using probe vehicles equipped with GPS devices [12] and transmitted to a central data collection point. PeMS is designed to accept data from any source as long as they are in electronic form. Regardless of the physical detection method, PeMS requires that the data are collected automatically.

2.1.4 Calculation of performance measures

In this section, we briefly explain how PeMS performance measures are computed from measurements of flow, occupancy, and speed.

Let $x_1, \ldots, x_n$ be locations of n detectors on a directional freeway, such as Interstate 5-Northbound in Los Angeles. They naturally segment the freeway into n parts, where the ith segment is the part of the freeway that is closest to the ith detector. For $1 < i < n$, the boundaries of segment i are $\frac{1}{2}(x_{i-1} + x_i)$ and $\frac{1}{2}(x_i + x_{i+1})$, and its length is

$$l_i = \frac{1}{2}(x_{i+1} - x_{i-1}).$$

The lengths of the first and last segments are defined as

$$\begin{aligned} l_1 &= \frac{1}{2}(x_2 - x_1), \\ l_n &= \frac{1}{2}(x_n - x_{n-1}). \end{aligned}$$

With a sample period of 30 seconds, at each sample time t, each detector i reports the vehicle count $Q_i(t)$, average occupancy $K_i(t)$, and average speed $V_i(t)$ for the sampling period ending at t. Actually, in most locations in California, speed is not directly measured but must be computed from Q and K by an algorithm described in Chapter 6. Chapter 4 gives a more detailed description of the entire data flow.

PeMS converts 30-second samples to 5-minute aggregates, and uses the 5-minute values of flow and speed to compute performance measures. Let $\mathrm{VMT}_i(t)$ and $\mathrm{VHT}_i(t)$ be the quantities for the ith segment and tth 5-minute period,

$$\mathrm{VMT}_i(t) \stackrel{\text{def}}{=} Q_i(t) l_i, \tag{2.3}$$

$$\mathrm{VHT}_i(t) \stackrel{\text{def}}{=} \frac{Q_i(t) l_i}{V_i(t)}. \tag{2.4}$$

VMT and VHT of different segments and time periods are additive, so they can be combined to give values for entire freeways and freeway systems. They are also aggregated into hourly, daily, monthly, and annual quantities.

Travel time $T_i(t)$ on a segment is simply

$$T_i(t) \stackrel{\text{def}}{=} \frac{l_i}{V_i(t)}. \tag{2.5}$$

Travel time of a long route, however, is not exactly the sum of $T_i(t)$ along that route, because the actual travel times of the segments depend on the travel times of the previous segments. We present an algorithm to find route travel times in Chapter 3.

S and D are computed from VHT and VMT as in (2.2) and (2.1).

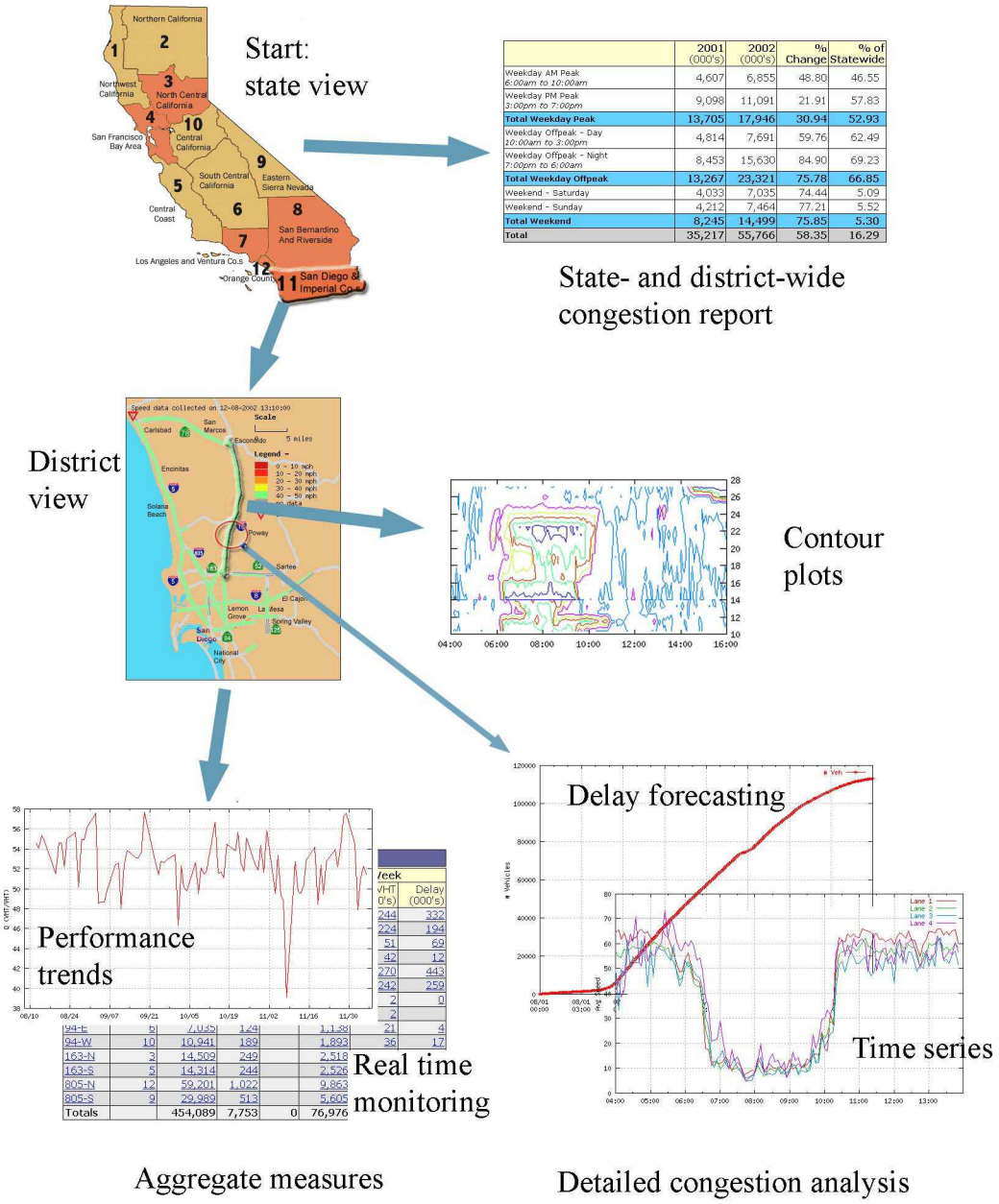
2.1.5 Using PeMS

PeMS is designed as a tool for traffic planners, policy makers, and engineers. It also has applications for detector maintenance and traveler information. Users access PeMS primarily through a Web interface. For example, the Caltrans director can open up PeMS on his PC and see state-wide performance figures. From that point, he can drill down to compare congestion in each district. Many levels of spatial and temporal aggregation are available, as well different ways of visualization. The navigation of PeMS is illustrated in Figure 2.1.

Caltrans is organized into 12 districts, each of which collects its own data. For example, San Francisco Bay Area and Los Angeles are two of the districts. PeMS is also organized by the same districts. The current freeway speeds and incidents are displayed on the front page of each district, as well as highlights showing the most congested locations.

District-wide performance is shown in plots of VMT, VHT, delay, etc. They can be broken down by freeway and time of day, so a transportation planner can use them to recognize trends in and respond to them.

Traffic engineers use PeMS to diagnose specific congestion problems, such as a persistent bottleneck or an area with high rates of incidents. PeMS provides many data visualization tools, including contour plots and plots of speed, flow, and occupancy in time and space. Its applications forecast the effect of lane closures and the incidents, and estimate the benefit of ramp metering policies. There are also online diagnostics tools for detector maintenance and an automated traveler information service. The rest of this chapter describes PeMS's Web-based tools.



	2001 (000's)	2002 (000's)	% Change	% of Statewide
Weekday AM Peak 6:00am to 10:00am	4,607	6,855	48.80	46.55
Weekday PM Peak 3:00pm to 7:00pm	9,098	11,091	21.91	57.83
Total Weekday Peak	13,705	17,946	30.94	52.93
Weekday Offpeak - Day 10:00am to 3:00pm	4,814	7,691	59.76	62.49
Weekday Offpeak - Night 7:00pm to 6:00am	8,453	15,630	84.90	69.23
Total Weekday Offpeak	13,267	23,321	75.78	66.85
Weekend - Saturday	4,033	7,035	74.44	5.09
Weekend - Sunday	4,212	7,464	77.21	5.52
Total Weekend	8,245	14,499	75.85	5.30
Total	35,217	55,766	58.35	16.29

94-E	6	7,035	124		1,138	21	4
94-W	10	10,941	189		1,893	36	17
163-N	3	14,509	249		2,516		
163-S	5	14,314	244		2,526		
805-N	12	59,201	1,022		9,863		
805-S	9	29,989	513		5,605		
Totals		454,089	7,753	0	76,976		

Figure 2.1: PeMS navigation and some applications.

2.1.6 Long term trends

By tracking performance measures over time, we can gauge the growth in demand and congestion, and identify locations that need the most improvement. We can also compare the performance before and after congestion-improvement projects to evaluate their effectiveness.

The following are delay measurements from District 4 HICOMP reports [9]. District 4 is San Francisco Bay Area. In addition to district-wide measures, HICOMP also

Year	Daily delay (1,000 vh)	Change from previous (%)	Cost per day (1,000 dollars)
1998	112	24	1,249
1997	–	–	–
1996	90	31	841
1995	68.5	14	641
1994	60.4	-5	565

Table 2.2: Average total daily delay in Caltrans District 4, from HICOMP reports. No report was produced in 1997.

provides congestion measure by route and by county. The purpose of the HICOMP report is

> . . . evaluating freeway performance for the purpose of establishing priorities and directing resources towards the areas most impacted. This data may also be used to evaluate effectiveness of the various strategies used to reduce congestion by comparing congestion before and after their implementation. [9]

PeMS produces standard and customized reports from its on-line database. Figure 2.2 shows the PeMS HICOMP report interface and output for District 8. As Figure 2.2 shows, one can choose to view a number of quantities for each district. Figure 2.2 shows a detailed analysis of congestion by time of day and weekday/weekend.

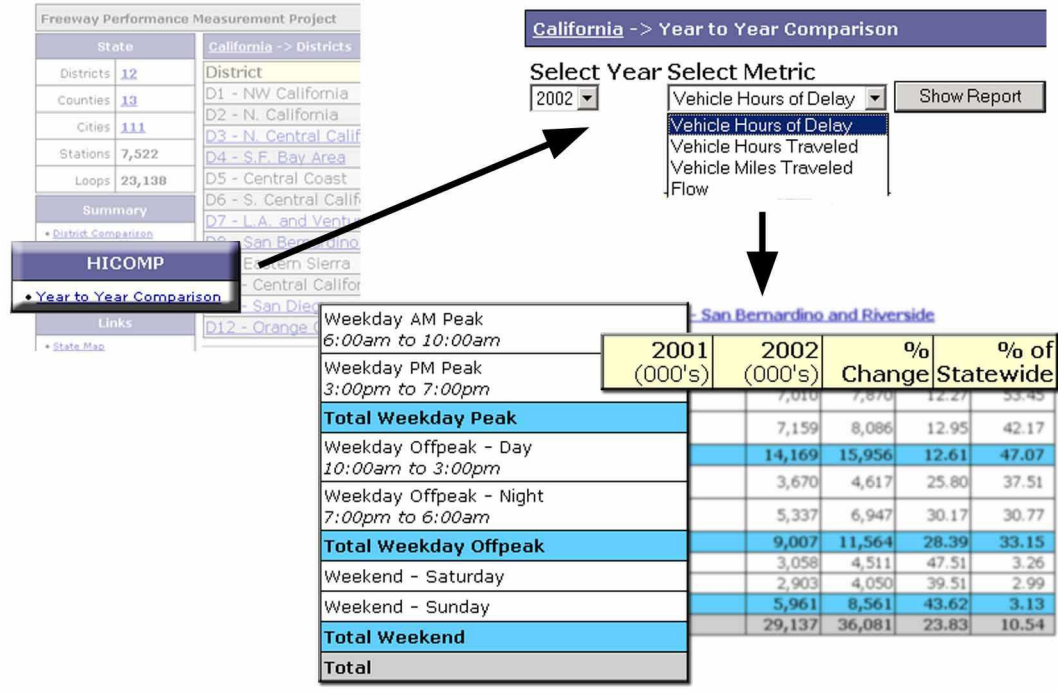


Figure 2.2: PeMS HICOMP report interface.

PeMS replaces the need for HICOMP because of its better accuracy and lower cost. It is accurate because it uses much more data. As we have seen, current data collection is manual and sporadic. But freeway congestion is highly variable. Figure 2.3 shows the delay variation on mid-week weekdays in 2001-2002. This figure shows that the daily delay varies between 0.5 million hours and 2 million vehicle-hours. Current Caltrans performance reports (HICOMP,AADT) use only a few days of data to estimate annual figures. Therefore, their results contain large errors.

2.1.7 Measuring freeway efficiency

Aggregate measures can be used to evaluate the relative performance of locations and time periods. Is the congestion worse now that it was last month? Is the performance of

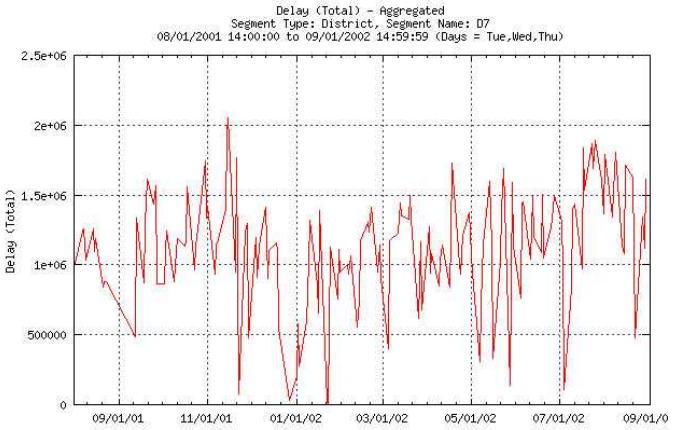


Figure 2.3: Variation in daily delay on mid-week weekdays, Los Angeles, 2001-2002.

one freeway worse than another? For example, some people believed that traffic congestion in September is worse than that in July and August [13]. PeMS data can be used to measure this phenomenon. Figure 2.4 shows a plot of average daily morning peak hours S for each

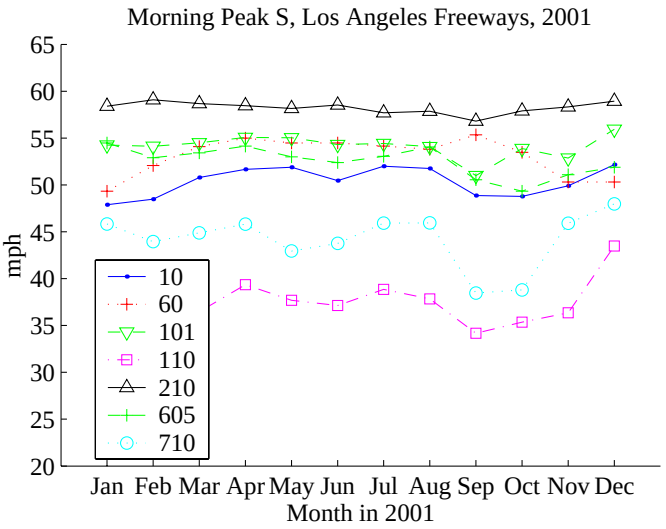


Figure 2.4: Average speed (S) on several Los Angeles freeways in 2001.

month of 2001, for several freeways. Peak morning hours are chosen to be between 5 and

10 AM. For 6 out of the 7 freeways studied, the average speed in September is indeed lower than it was in August. This information suggests that extra effort should be applied to improve traffic in September.

We use different measures in different types of comparisons. For example, VHT is additive and applies to a freeway segment, an entire freeway, and a region. It's an overall measure of the cost of travel, and can be readily translated into dollars. On the other hand, VHT can be misleading if used to compare two different freeways, because a longer freeway is likely to have more VHT than a shorter one, but it also serves more people. S is a better measure for this purpose.

Figure 2.4 shows the monthly average S on several Los Angeles freeways in 2001. It shows that average peak hour speeds on Freeway 110 are consistently lower than those on other freeways. Traffic engineers can look at a comparison like this one to determine the locations of the worse congestion and investigate them further to find a remedy. We discuss some the PeMS tools that help the engineer to solve specific problems in Section 2.2.

2.1.8 Incident statistics

Figure 2.5 shows the distribution of collisions in Los Angeles in July 2002. This figure shows that I-10 had the most collisions, followed by I-405 and I-5. This is probably because these are the longest freeways in Los Angeles.

Traffic engineers can use incident statistics to detect locations that have high collision rates by plotting the spatial distribution of collisions on each freeway. For example, Interstate 210 West (I-210W) is a busy freeway that carries traffic to and from downtown Los Angeles. Figure 2.6 shows a map of its vicinity. Using PeMS's incident database, we

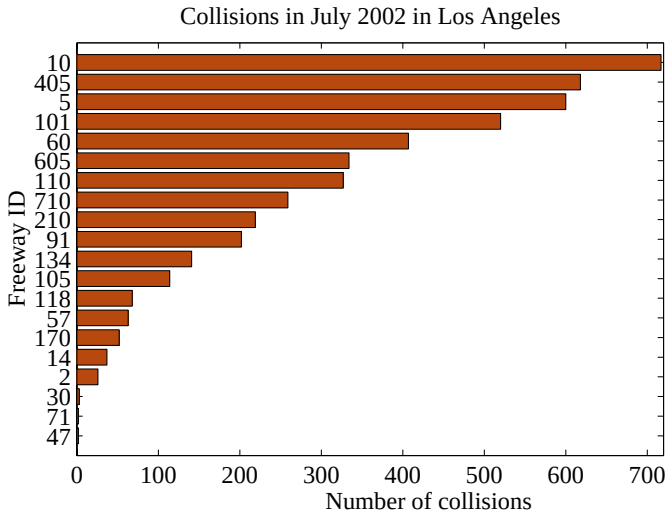


Figure 2.5: Collisions in Los Angeles in July 2002. Data source: CHP website [1].

Figure 2.6: Map of I-210W's vicinity.

plot the monthly rate of collisions on I-210W during March through August, 2002. Figure 2.7 shows that collisions on I-210W are concentrated near postmile 26.5, and also postmile 34 to a lesser extent. The higher incident rates at these locations could be a result of heavy traffic, especially since they are near two freeway interchanges with SR-134 and I-605. We

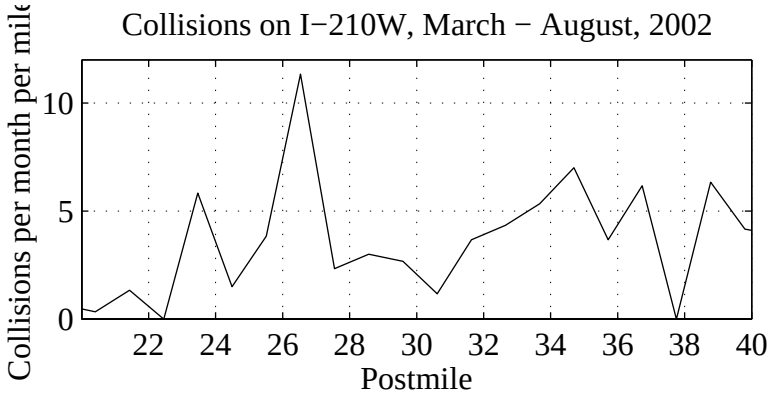


Figure 2.7: Collisions on I-210W in Los Angles between March and August, 2002.

investigate whether the peaks in Figure 2.7 correspond to the VHT at these locations. Since VHT of a location measures the amount of time vehicles spend there, it is an indicator of the "opportunity" to get into a collision. Figure 2.8 shows the average daily VHT on I-210W during the same period. VHT is highest at postmile 26, corresponding to the peak

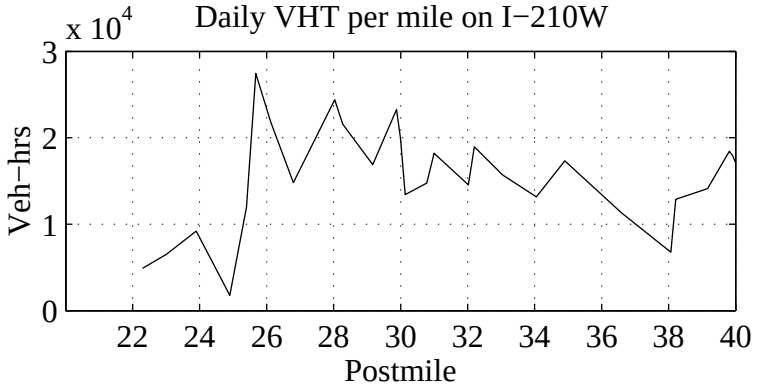


Figure 2.8: Average daily VHT on I-210W during March - August, 2002.

in collision rates in Figure 2.7. However, the peaks in VHT are not as sharp as those in collision rates. This example suggests that collisions in Figure 2.7 are caused not only by heavy traffic, but also by vehicles merging near freeway interchanges.

2.1.9 Real time monitoring

While the traffic planner is interested in the long term trends in traffic performance, the traffic operator is interested in the real time performance of the system. The real time data collection and on-line accessibility of PeMS means that current delay, VMT, VHT, and S are available to the traffic operator.

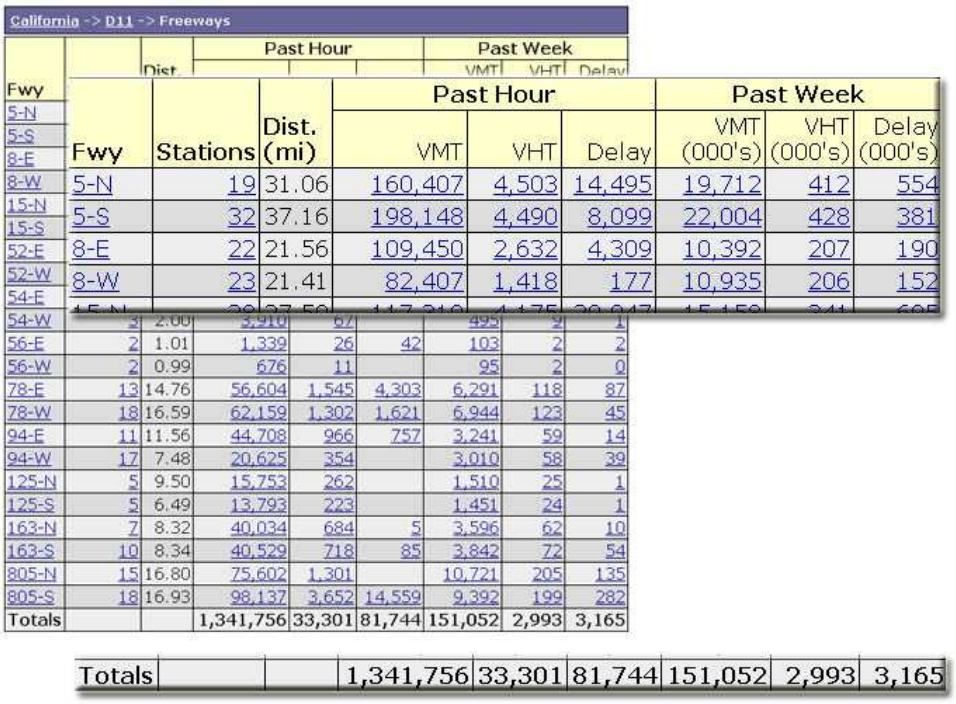
In each district, the locations with the worst current congestion are listed in a table that is refreshed every 5 minutes, shown in Figure 2.9 under "Slowest Speeds." This

Figure 2.9: District highlights.

is a screen shot of the PeMS page on Saturday, 11/23/2002, at 4 pm. Some of the locations in this list may be chronically congested at these times, while others may be encountering non-recurrent congestion that requires special attention, such as accidents or special events. These locations can be investigated further using PeMS's detailed plotting tools, discussed

in Section 2.2.

In addition to the most congested spots, the PeMS site displays a list of all the freeways in each district and their current conditions in the past hour and past month, shown in Figure 2.10. This view allows the operator to compare a freeway's current performance



California -> D11 -> Freeways

Fwy	Dist.		Past Hour			Past Week		
	Fwy	Stations	Dist. (mi)	VMT	VHT	Delay	VMT (000's)	VHT (000's)	Delay (000's)
5-N	5-N	19	31.06	160,407	4,503	14,495	19,712	412	554
5-S	5-S	32	37.16	198,148	4,490	8,099	22,004	428	381
8-E	8-E	22	21.56	109,450	2,632	4,309	10,392	207	190
8-W	8-W	23	21.41	82,407	1,418	177	10,935	206	152
15-N									
15-S									
52-E									
52-W									
54-E									
54-W	3	2.00	3,910	67		495	9	1	
56-E	2	1.01	1,339	26	42	103	2	2	
56-W	2	0.99	676	11		95	2	0	
78-E	13	14.76	56,604	1,545	4,303	6,291	118	87	
78-W	18	16.59	62,159	1,302	1,621	6,944	123	45	
94-E	11	11.56	44,708	966	757	3,241	59	14	
94-W	17	7.48	20,625	354		3,010	58	39	
125-N	5	9.50	15,753	262		1,510	25	1	
125-S	5	6.49	13,793	223		1,451	24	1	
163-N	7	8.32	40,034	684	5	3,596	62	10	
163-S	10	8.34	40,529	718	85	3,842	72	54	
805-N	15	16.80	75,602	1,301		10,721	205	135	
805-S	18	16.93	98,137	3,652	14,559	9,392	199	282	
Totals			1,341,756	33,301	81,744	151,052	2,993	3,165	

Totals			1,341,756	33,301	81,744	151,052	2,993	3,165

Figure 2.10: PeMS District summary by freeway.

with its performance in the past week and spot those that are behaving abnormally.

2.1.10 Service quality measured in travel time

VMT and VHT measure the aggregate output and input to the transportation system, but the service quality experienced by the customers, i.e. the drivers, is best represented in terms of travel time. While the VHT is itself a measure of average travel

time, it does not tell the whole story because it doesn't capture the variation in individual travel times. The variation in travel time is an important measure of the true cost to the driver.

Think of the case where you need to arrive at the airport to catch a plane by 7 pm on a weekday. How much time do you budget for the trip from your office to the airport? Note that this time must be longer than the average travel time – you need to put in a cushion to make sure you will catch the plane under most circumstances. In fact, you may decide to leave with enough time such that you will arrive on time with 90% probability. Then, the actual time cost of the trip is the 90th percentile travel time.

PeMS computes travel time statistics on all freeway segments for departure times at 5-minute intervals. Section 3.2 gives The details of this calculation. Figure 2.11 shows the travel time variability on a 20-mile route along I-5N in Los Angeles. At 5:00 am, the

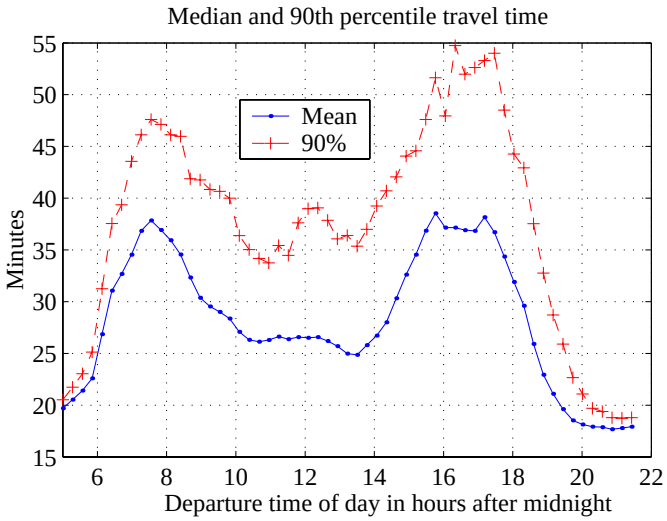


Figure 2.11: Travel time and 90th percentile travel time for each departure time on I-5N, between postmiles 0 and 20, in Los Angeles.

expected travel time is 20 minutes. The travel time is much higher during the morning

and evening peak hours. The average travel time is represented by the solid line, while the dashed line shows the 90th percentile travel time for the corresponding departure times. The 90% travel time is about 10 minutes higher than the average. This means that in addition to congestion delay, about 10 minutes per trip is lost due to scheduling for uncertainties.

2.2 Detailed analysis

PeMS has many data visualization tools that help traffic engineers diagnose problems such as chronically congested locations or locations with high incident rates. To unlock the information contained in PeMS's vast database, it is often necessary to visualize data from various perspectives and at different aggregation levels. For example, we can plot the delay on a freeway across a year and observe its trends, or we can plot the speed at one location over several hours to determine the exact onset of congestion. Table 2.3 shows a list of the types of plots available in PeMS. Figure 2.12 illustrates the levels of spatial and

Plot type	Example
Geographical	Real time speeds
Aggregate measures	Delay on a freeway over one year
Contour plots	Contour of speeds on a freeway over several hours, showing the progression of congestion in space and time
Quantity versus space	Speed at locations along a freeway at one instant in time, showing regions of high and low speeds.
Quantity versus time	Flow versus time for a day, showing periods of high and low flows
Quantities	Flow versus occupancy, the fundamental diagram for a detector location

Table 2.3: Types of plots in PeMS.

temporal aggregation of the different plots.

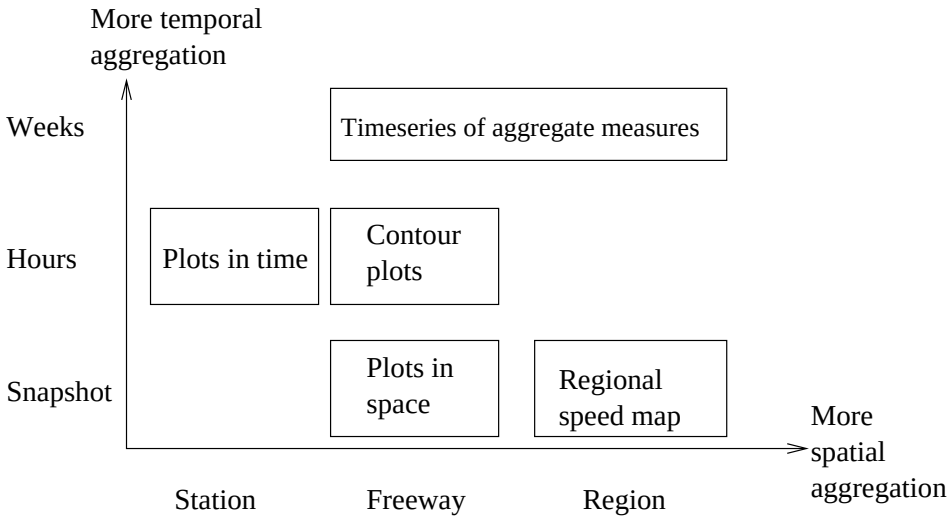


Figure 2.12: Types of plots and their level of aggregation.

2.2.1 Bottleneck analysis

It's best to use an example to understand PeMS's plots. The usual starting place is a plot of aggregates, such as a plot of S over several months. Figure 2.13 shows the daily S on one freeway, I-805 North, on weekdays between 7/15/2002 and 12/1/2002. The dotted line is the daily S and the solid line shows the trend, obtained by filtering with a window length of 7 days. The filtered S varies between 50 and 55 mph. In the beginning of September, for example, S dropped to about 50 mph. On 9/3/2002 through 9/5/2002, the average speed was a few miles per hour lower than they were in the past. Studying the cause of the delay may help us prevent similar congestion in the future.

We choose one of these days, 9/5/2002, for a more detailed analysis on the cause of congestion, starting with a contour plot of its speeds. The contour plot is quite informative because it shows the progression of congestion in time and distance. Figure 2.14 shows a large congestion region, marked by the contours of equal speeds. In this figure, travel is

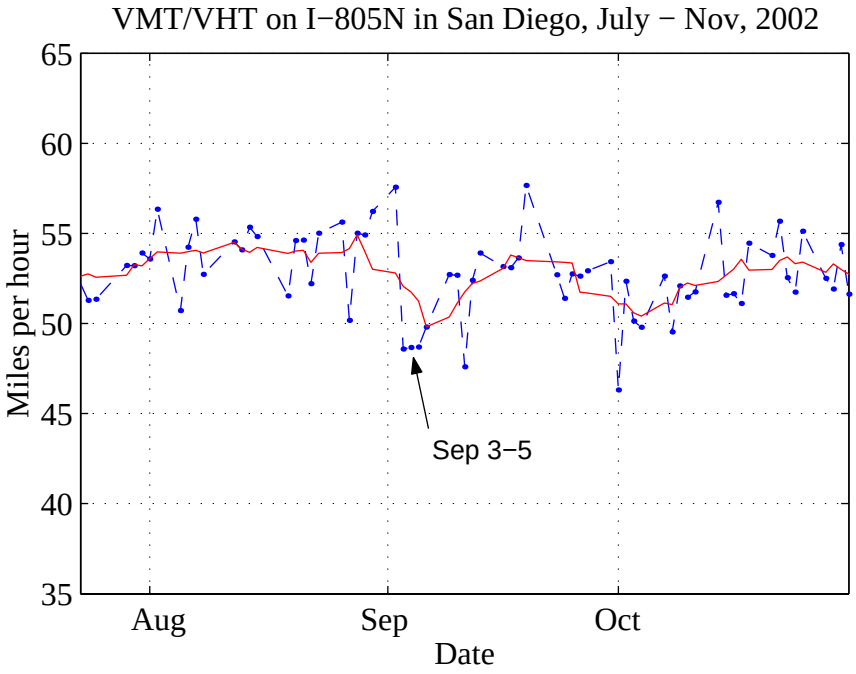


Figure 2.13: Trend in S over several months.

in the direction of increasing postmile. There appears to be a bottleneck at postmile 24. Upstream of mile 24, speeds are low between 6:30 and 9:30; downstream speeds remain high throughout the period. Congestion starts out at mile 24, and reaches mile 14 by 7:45.

While the contour plot shows regions of congestion, it's harder to see the depth of congestion from the contours alone. A more quantitative view of congestion is found in a plot of speeds at locations on the freeway at a sample time. Figure 2.15 shows the speeds on I805N at 8 am, where traffic flows from left to right. Here, the congested region is shown clearly as between miles 14 and 24, where the speeds are below 25 mph.

Contour plots of several other weekdays confirm that the region upstream of postmile 24 is chronically congested. Therefore, the congestion on 9/5/2002 is probably re-

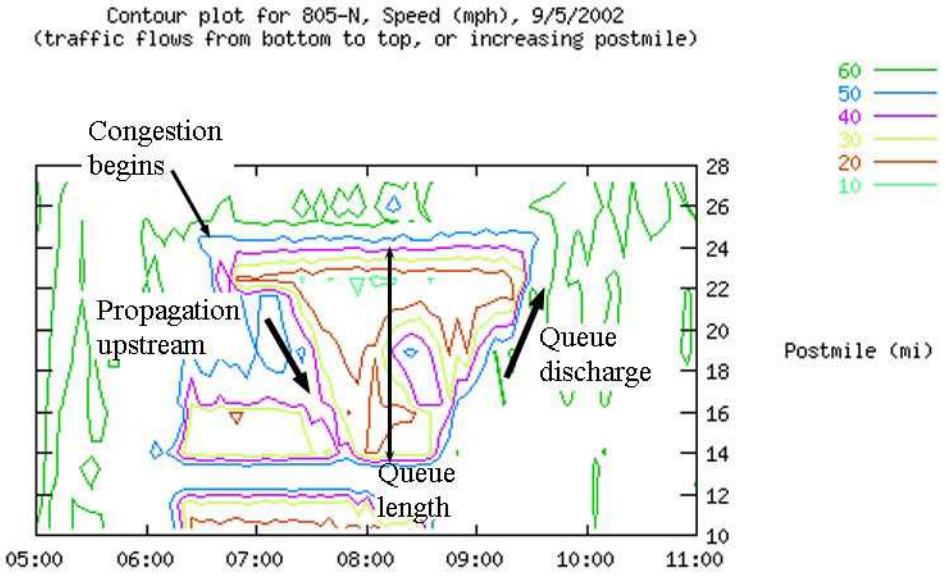


Figure 2.14: Contour plot of I-805N on 9/5/2002 between 5:00 and 11:00 AM.

current and not caused by an incident. We plot the speeds measured at the upstream and downstream detectors for several hours on 9/5/2002. Figures 2.16 and 2.17 show a sustained drop in speed between postmiles 24.41 and 22.48, indicating that a queue has formed between these locations. The downstream detector (at postmile 24.41) is measuring the discharge of the queue. This queue is probably caused by the merging traffic from SR-52, which crosses I-805 between these detectors. Therefore, Figure 2.16 shows the speed inside the queue, which are very low during the congested period. At the downstream, shown in Figure 2.17, the traffic has gotten past the bottleneck, and is traveling at a much higher speed than at the upstream. There is still a speed drop at postmile 24.41 during the congested period of between 6:30 and 9:30 AM compared to other periods at the same location. However, the speed drop is not severe, and speeds during the peak period remained above 45 mph. This speed drop may be because of vehicles accelerating out of the queue.

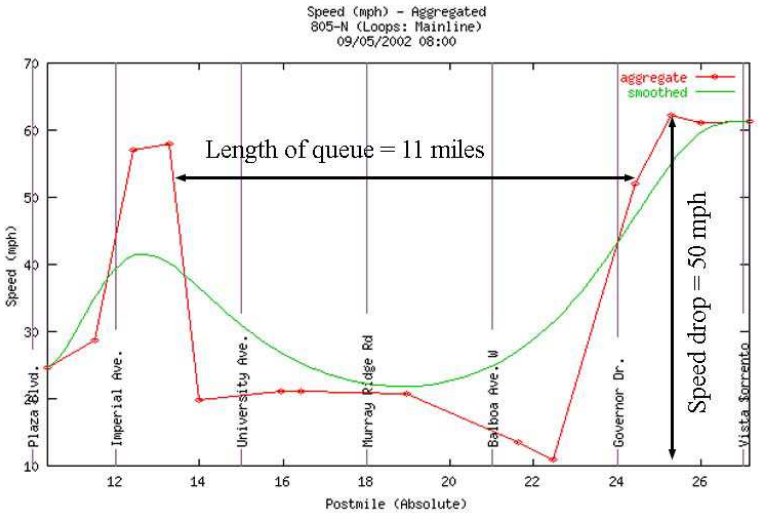


Figure 2.15: Plot of speeds on I-805N at 8:00 AM on 9/5/2002. Traffic flows from left to right.

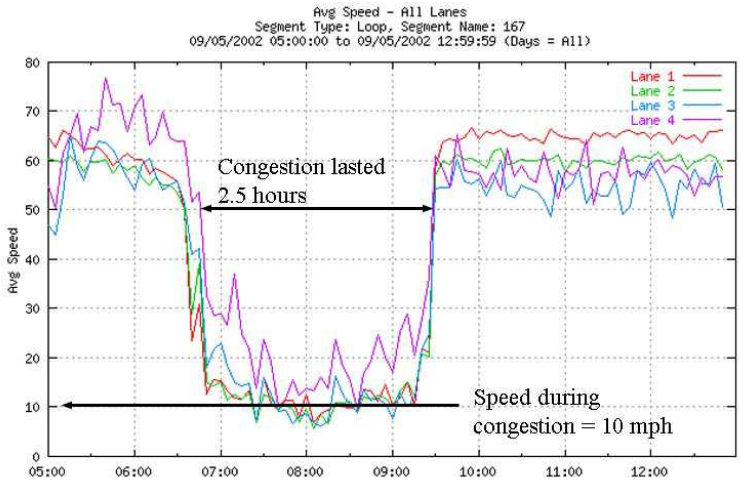


Figure 2.16: Flow upstream of the bottleneck, inside queue at postmile 22.48.

The congestion mechanism can also be studied using plots of flow rate versus time. We can use them to observe the change in flow as the bottleneck formed and dissipated. Figure 2.18 shows the flow rate upstream of the bottleneck. At 6:30, the flow rate was 8000 vph. By 8:00, it dropped to 4000 vph, while the speeds at this time has dropped from

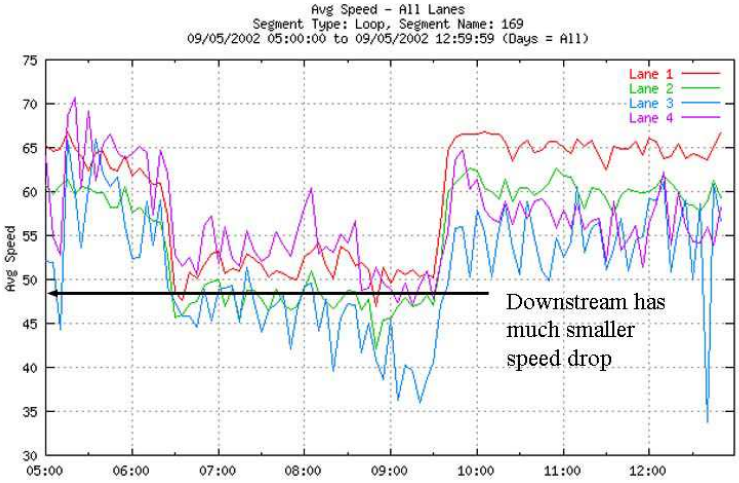


Figure 2.17: Speed downstream of the bottleneck, queue discharge flow at postmile 24.41.

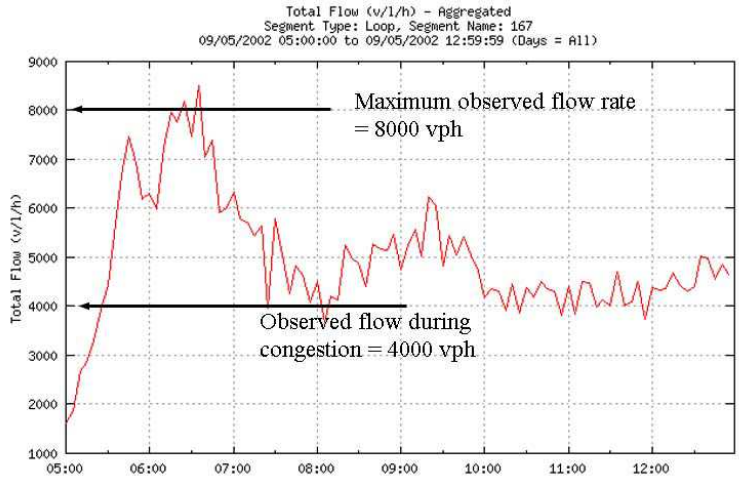


Figure 2.18: Speed upstream of the bottleneck, inside queue.

60 mph to about 10 mph. Why did the flow rate drop so much? During the same period, the flow rate downstream at mile 24 remained nearly constant, at about 8500 vph. The difference in flow of the upstream and downstream must be coming from another source, probably SR-52. However, there are no detectors on SR-52 at this location so this is not

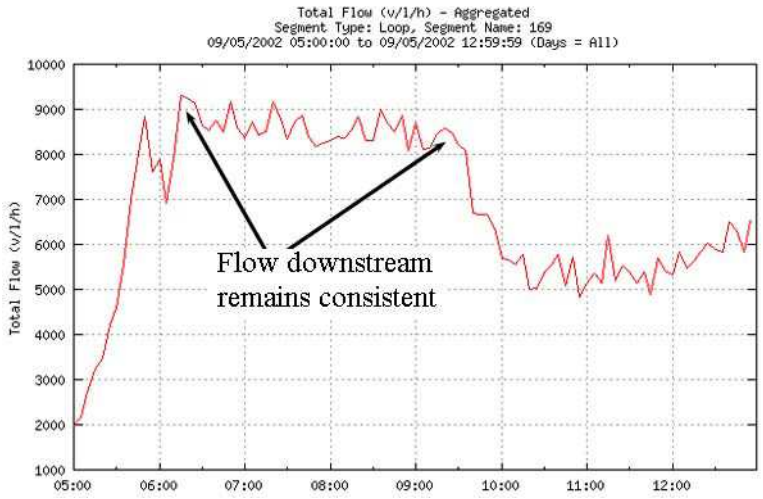


Figure 2.19: Flow downstream of the bottleneck, queue discharge flow.

verified.

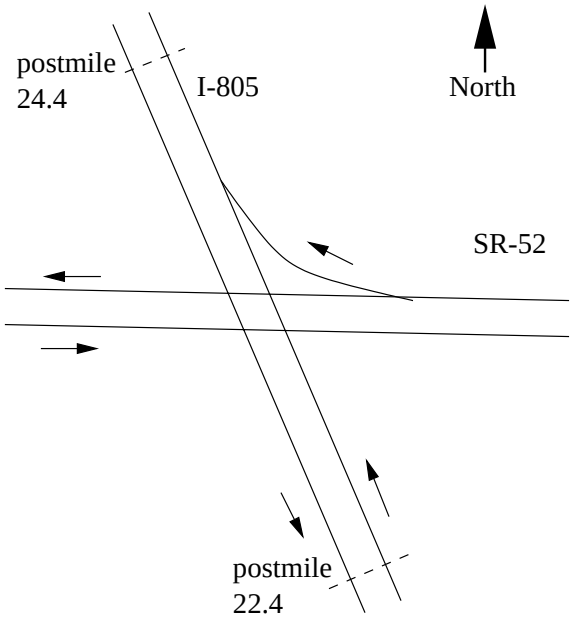


Figure 2.20: I-805 and SR-52 exchange between postmiles 22.4 and 24.4 on I-805.

This exercise suggests that congestion at this location is caused by the merging

traffic from another freeway. When the cause of congestion is known, we can begin to think of ways to improve the situation, by increasing the downstream capacity or using ramp metering to facilitate better merging. The above analysis shows how progressively more detailed plots can be used to investigate freeway locations.

2.2.2 Incident analysis

PeMS plots can also be used to analyze incidents. We take an incident from the database whose details are in Table 2.4. This incident is serious because it lasts for 3 hours

Incident location:	SB I15 at Pomerado Rd
Incident type:	Traffic collision – ambulance responding
Start time:	8-1-2002 14:19
End time:	8-1-2002 17:10

Table 2.4: Incident detail on I-15S on 8/1/2002 in San Diego.

and required an ambulance. What was its effect on the traffic at this location? A traffic analyst at the TMC can use PeMS's plotting tools to get the answer.

This incident occurred on a Thursday, but before the rush hour traffic. In fact, at this location the congestion is usually light. Pomerado Road is located at mile 26 on I15. There was no congestion at this location on the previous day, according the contour plot of speeds on 7/31/2002 shown in Figure 2.21. On 8/1/2002, however, this location is engulfed in deep congestion beginning at 14:20, shown in Figure 2.22.

This incident severely reduced flow rate. Figure 2.23 shows that the flow rate dropped from 6500 to 2500 vph when the incident occurred. The horizontal axis marks the hours of day, and the vertical marks flow rate in vph. Speed at this location also dropped

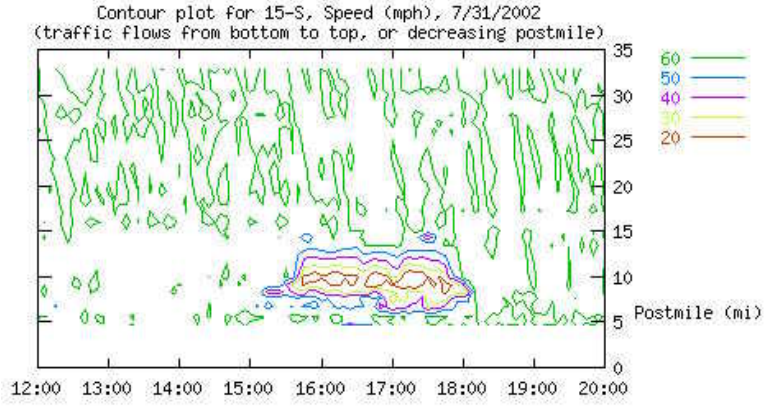


Figure 2.21: Contour plot of speed on 7/31/2002 on I-15S.

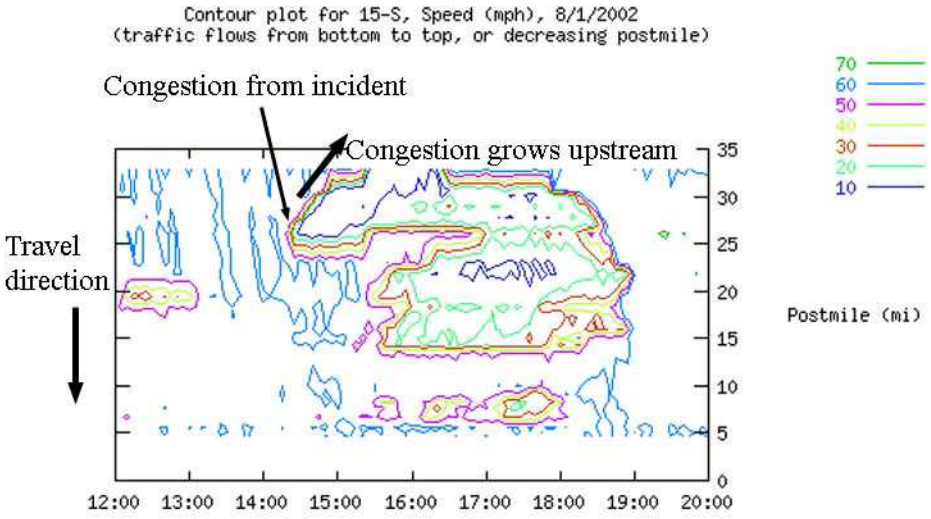


Figure 2.22: Contour plot on 8/1/2002 on I-15S shows congestion at mile 26, between 14:19 and 17:10.

from about 60 mph to below 20 mph for more than an hour, starting at 14:20. Figure 2.24 shows the lane speeds on this afternoon. Between 17:00 and 18:30, speed dropped again, this time due to recurrent commuter traffic.

Figure 2.25 shows the speed profile on 8/1/2002 at 16:00. Speeds are low for up to 7 miles upstream, between postmiles 26 and 33 (traffic flows from right to left), indicating

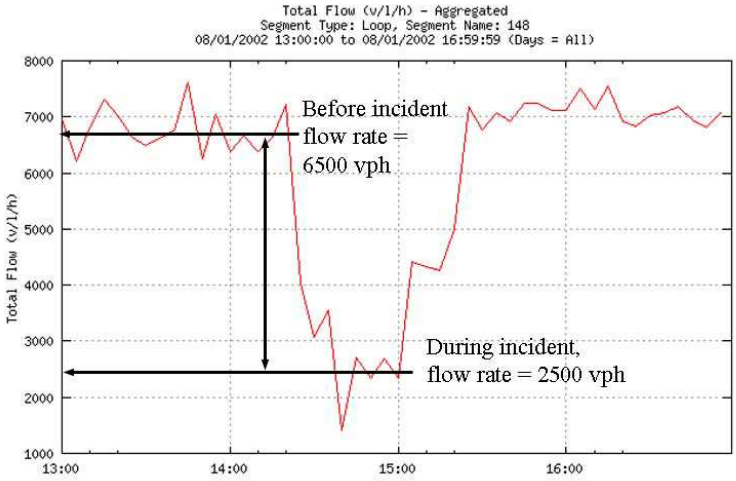


Figure 2.23: Total flow rate of location just upstream of incident at postmile 26.

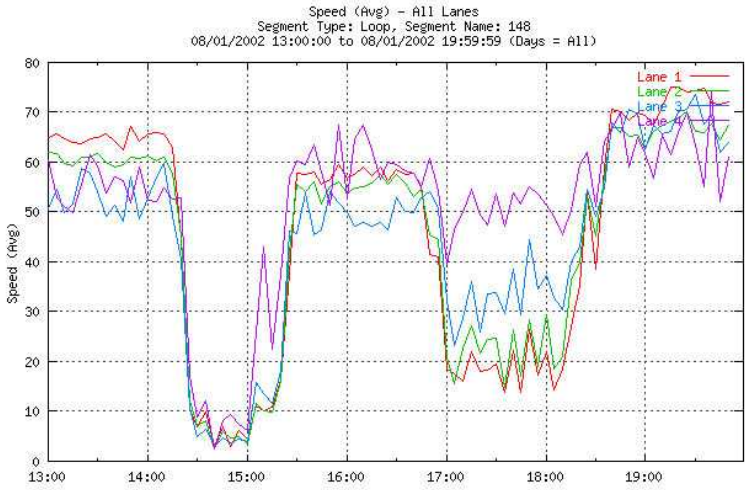


Figure 2.24: Speed dropped at 14:20 because of incident.

that the effect of the incident was felt far upstream.

This incident also restricted the flow rate along the freeway. Figures 2.26 and 2.27 show the freeway flow rates at the same time of day on 8/1/2002 and the same day the following week. The flow rate at Pomerado Rd (postmile 26) was 5500vph on 8/8/2002; but

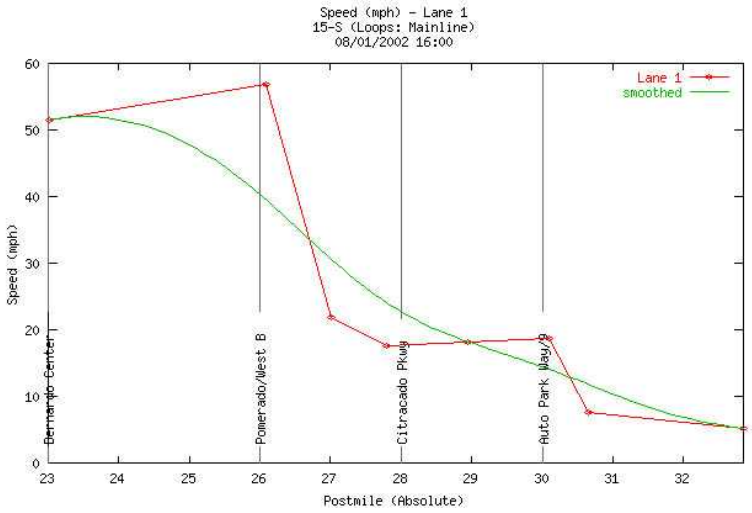


Figure 2.25: Speed profile on route.

on 8/1/2002, because of the incident, it was only 3000vph. In fact, flow was restricted to below 3000vph for 5 miles upstream of the bottleneck, whereas on 8/8/2002, the achieved flow rates were almost twice as high. At the same time, the downstream flow rates are much lower than what they were on the non-incident day. This incident indeed had a large impact on the traffic.

Finally, the quantitative increase in delay caused by the incident is seen clearly on a plot of delay versus time for several months. Figure 2.28 shows that this incident caused much high delays compared to historical values.

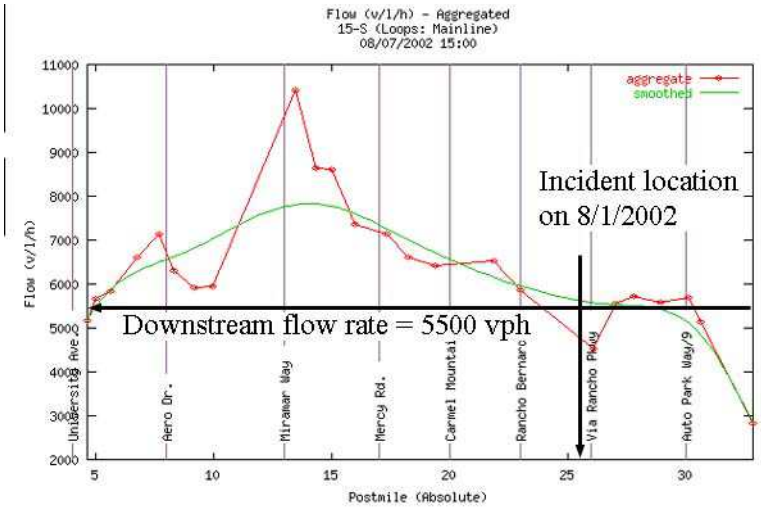


Figure 2.26: Flow rates on a normal day at 3 pm.

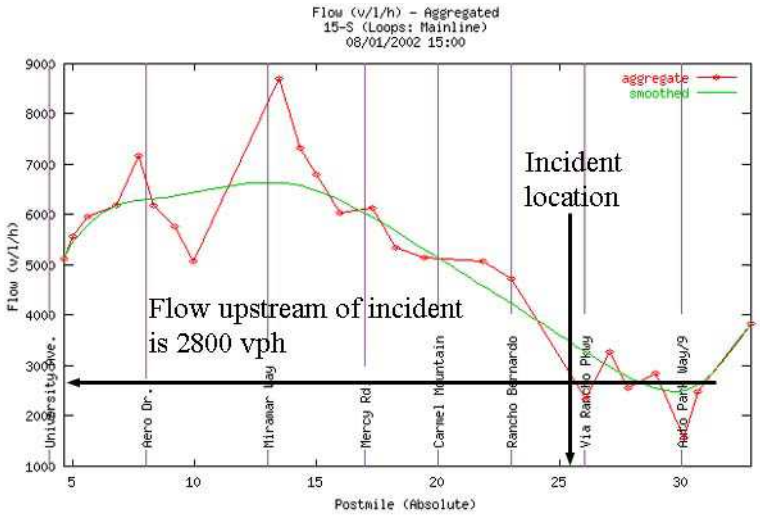


Figure 2.27: Flow rates on day of incident. They are much lower than usual.

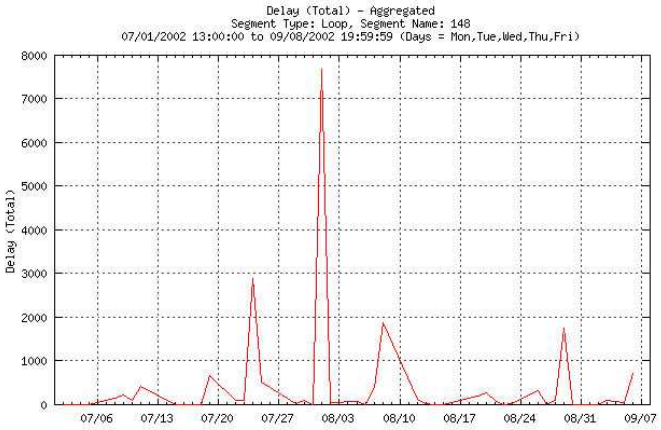


Figure 2.28: Delay over several months shows the incident on 8/1/2002 caused abnormally high delays at this location.

2.3 Forecasting tools

2.3.1 Delay forecasting of accidents and lane closures

By observing the impacts of incidents, we can predict the effect of similar incidents and benefits of quickly clearing away an incident. The incident in Figure 2.22 reduced the flow rate for 40 minutes, and caused 7500 vehicle hours of delay on 8/1/2002. What would have happened if a similar incident also occurred on 8/8/2002? How about if the time to clear was reduced by 10 minutes, how much delay can we save? Similar delays might result if the lanes were closed for construction as well as incidents. PeMS provides the Capacity Analysis tool that calculates the delay resulting from lane closures of a specified number of lanes and length of time. This is a useful tool to diagnose the potential effects of traffic accidents and construction projects.

The lane closure scenario can be described by a queueing system. Suppose traffic is free-flowing before time t_0, at a rate of r_d vehicles per hour. At t_0, an accident blocks

part of the road, forming a bottleneck at x. The departure rate becomes r_b, while traffic is still arriving with the demand rate of $r_d > r_b$. Therefore, vehicles build up behind the bottleneck in a queue, see Figure 2.29. This queue has $n_q(t)$ vehicles at time t and grows at the rate of $r_d - r_b$. The blockage is cleared at t_1. The departure flow rate increases to the freeway capacity at this point, r_c, where $r_c > r_d$. Since the departure rate is higher than the arrival, the queue discharges until it disappears at t_2.

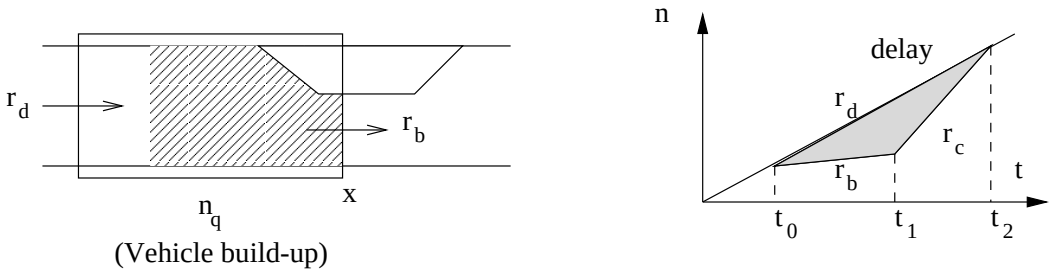


Figure 2.29: Queueing created by lane closure.

The right side of Figure 2.29 shows the cumulative vehicle count at location x, where the gray area represents the total delay. During the incident on 8/1/2002, Figure 2.23 shows that $r_b \approx 2500$ vph; the discharge rate after clearance (15:20 PM) was $r_c = 7000$ vph. The demand rate before the incident was about $r_d \approx 6500$ vph.

The Capacity Analysis tool simulates the delay resulting from lane closures of a specified severity and duration. To simulate the effect of a certain type of incident on a Monday, for example, one can use this tool to simulate the effect of the incident on a similar Monday for which we have historical flow data. We demonstrate how this tool is used to simulate the delay if an incident similar to this one happened on 8/8/2002 at the same time of the day.

Figure 2.30 shows the interface of the lane closure tool. The user enters the day and time of the analysis, the incident severity and duration, and lane capacities r_b and r_c. The demand r_d is taken to be the actual flow rates at the location.

Figure 2.30: Capacity analysis tool interface.

Figure 2.31 shows the result of the simulation, which is a plot of cumulative vehicle counts of the demand and the simulated discharge flow, similar to Figure 2.29. The cumulative demands from actual measurement are shown by the circle-line. The simulated impact of the incident is shown by the plain line. Between 14:20 and 15:00, the simulated flow is restricted to the low rate of 2500 vehicles per hour. After the incident is cleared at 15:00, the queued vehicles discharge at the capacity rate of the link. The point where the simulated flow meets the actual flow at about 18:10 represents the time it takes for the link to become uncongested. The area between the real cumulative flow and the simulated flow is the simulated impact of the incident, measured in vehicle-hours. This figure shows that the specified incident on this day would have created congestion lasting four hours. The results are also summarized in Figure 2.32, which shows an estimated delay of 5764 vehicle-hours.

Figures 2.31 and 2.32 show the effect of an incident with a duration of 40 minutes.

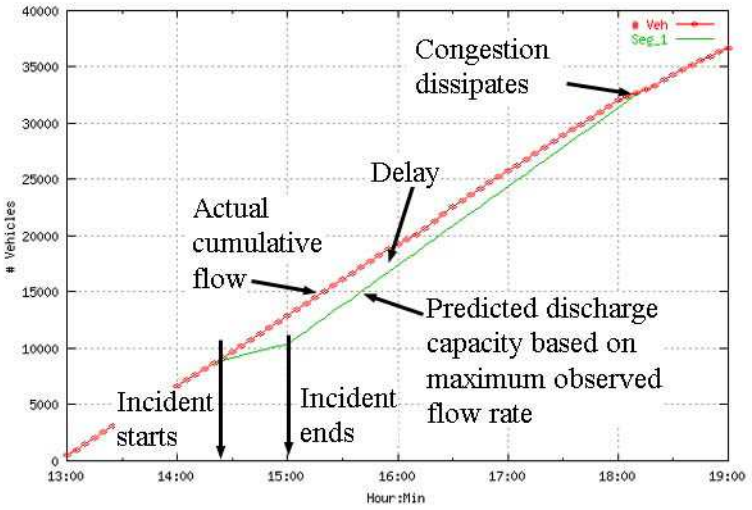


Figure 2.31: Capacity analysis – predicted effect of incident on 8/8/2002

Seg #	Incident Start Time	Delay Start Time	Delay End Time	Segment Delay Duration (minutes)	Incident Capacity (veh/hr)	Max Flow During Segment (veh/hr)	Max Individual Delay (minutes)	Max Queue Length (miles)	Total Delay (veh-hrs)
1	08/08/2002 14:20:00	08/08/2002 14:20:00	08/08/2002 18:15:00	235	2500	7884	22.9 at 08/08/2002 14:40:00	2.99	5764.3
								Sum:	5764.3

Figure 2.32: Capacity analysis – predicted effect of incident on 8/8/2002

The forecasting tool allows us to ask, how much of the delay can be eliminated the incident can be cleared faster? When we input 30 minutes as the incident duration, the resulting delay becomes only 3379 vehicle-hours. Therefore, a 10 minute reduction in lane blockage resulted in a reduction of 1400 vehicle hours, suggesting that the quick clearance of incidents can greatly improve incident-caused delays.

The capacity analysis tool can also be used to predict the effect of lane closure for construction projects. In 2001-2002, Caltrans spent $4 billion in capital outlay projects, out of a total highway budget of $6.8 billion [6]. Therefore, major part of Caltrans' budget

Seg #	Incident Start Time	Delay Start Time	Delay End Time	Segment Delay Duration (minutes)	Incident Capacity (veh/hr)	Max Flow During Segment (veh/hr)	Max Individual Delay (minutes)	Max Queue Length (miles)	Total Delay (veh-hrs)
1	08/08/2002 14:20:00	08/08/2002 14:20:00	08/08/2002 17:50:00	210	2500	7884	16.8 at 08/08/2002 14:35:00	2.19	3379.9
								Sum:	3379.9

Figure 2.33: Capacity analysis – predicted effect of incident on 8/8/2002, if incident is cleared in 30 minutes instead of 40 minutes.

is devoted to maintaining and improving the condition of the roads. While construction is very expensive in labor and materials , it also introduces congestion and delay. When road closures are required for a construction project, such as the repavement of an existing freeway, there are often choices of how many lanes to close, and for how long at a time. This decision has an impact on the total construction cost as well as the total delay that results. For example, it may be more efficient to close all 4 lanes at a time than two at a time when repaving a 4-lane freeway, but closing all 4 lanes results in more delays [14]. The best lane closure strategy optimizes the overall cost in both construction and delay. While construction costs of various closure schedules can be straightforwardly estimated, estimating the delay requires knowledge of the traffic demand at the location for the times of the closures. Using the PeMS delay forecasting tool shown in Figure 2.30, we can predict the delay resulting from a given closure strategy. Therefore, this is a useful tool for optimizing lane closure schedules.

2.3.2 Ramp metering gain calculator

Similar to the lane closure tool, we can apply other models on the data to evaluate what-if situations. PeMS implements a simulator that predicts the potential delay reduction

of ramp metering. We know that the capacity of the freeways is higher under free flow conditions than congested. For more on this result, see Section 3.1 and [15],[16]. If free flow is maintained through ramp metering, a higher flow rate can be achieved than under the un-metered condition. The PeMS ramp metering tool calculates delay savings using real data.

Given a freeway segment containing a number of on ramps and off ramps, the ramp metering calculator estimates the delay savings in vehicle hours traveled during the specified period under ideal metering. This calculation uses actual flow rates as the demand, and applies metering rate that restricts ramp flow to maintain free flow on the mainlines and speeds at 60 mph. The observed dependencies between flow and occupancy are used to determine the desired occupancy level and metering rate.

Figure 2.34 shows the result of the metering simulation performed on the data from I-210W in Los Angeles in the AM peak hours of Jan 11, 2001. The vertical axis shows VHT in vehicle-hours. The top curve shows the actual VHT. The bottom line shows the minimum VHT, which is the resulting VHT if every vehicle was allowed to travel at 60 mph. The middle curve is the delay under ideal metering. Ideal metering means we meter at the on-ramps so the flow on the freeway is always free flow. The difference between ideal metering and the minimum VHT represent the time spent waiting on the ramps, and represents the delay due to excess demand. This plot shows that about 80% of the total delay is caused by inefficient operation, and only about 20% is from demand in excess of capacity.

The ramp metering tool can be applied to any freeway stretch to predict the

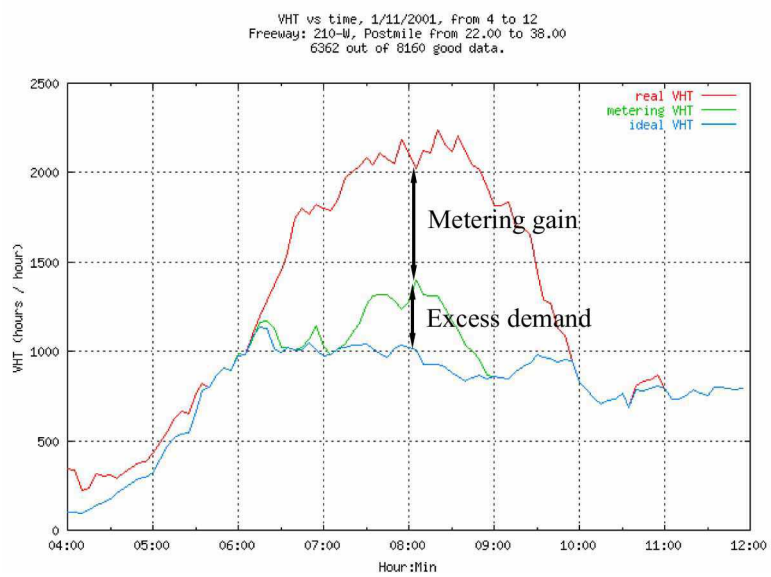


Figure 2.34: Reduction in delay from ideal metering, a simulation on real data.

potential gain from ramp metering at different locations. We can use the results to determine the places where metering will realize the highest delay savings.

2.4 Detector diagnostics

PeMS depends on accurate and comprehensive data. Therefore, an important part of the system is data quality monitoring and detector diagnostics. There are 23,138 loop detectors[1] in the PeMS database all over California. They operate in the harsh outdoor environment of sun, rain, and cold. Many detectors fail and produce bad or missing measurements. PeMS generates reports indicating detector health. Figure 2.35 shows the status of each freeway in Los Angeles. The detectors are placed in four categories listed in Table 2.5. These reports allow detector maintenance persons to identify locations of worst quality

[1]There are 7,522 *stations*, each station has several lanes, including mainline, HOV, and ramps.

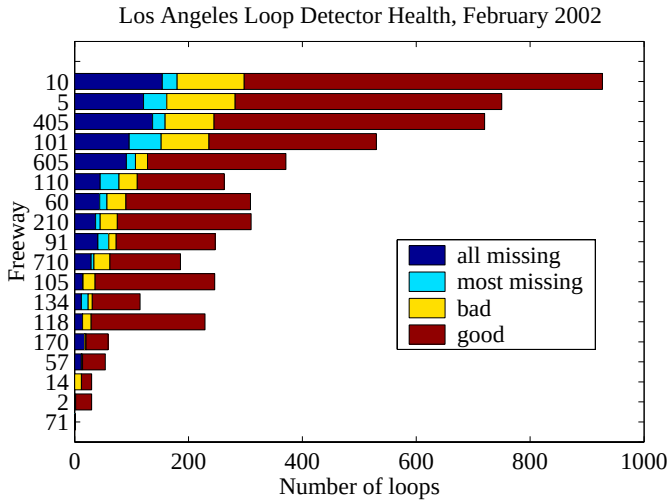


Figure 2.35: Loop quality by freeway

Status	*% of total*
Always missing data	16%
Missing most data	5%
Have most data, but most are bad	12%
Mostly good	67%

Table 2.5: Categories in Figure 2.35.

and dedicate resources to correct the problems. A detailed diagnosis of each detector can be found in the PeMS database. PeMS diagnoses detectors daily and generates reports that list error conditions.

PeMS gives the repair person advanced knowledge of the detectors before he visits them physically. Currently, loop maintenance is tedious and not very effective. The Caltrans Detector Fitness Program("DFP"), for example, is a program that inspects loop detectors. In DFP, crews of maintenance workers drive to each loop location on a freeway and perform a set of electrical measurements on the equipment, notes malfunctions, and types in a field

report. They fix certain types of malfunctions. However, the crew never looks at the loop data, either before or after the visit, so they don't know if the detector has been fixed, or if it was broken in the first place. We found that some DFP field measurements are not informative about actual data quality. Using PeMS's diagnostics tools, instead of visiting every single loop location, the repair person needs only to visit malfunctioning ones. In Los Angeles, a crew of two spends 1-2 hours at each location. Their time would be better spent if it's focused on malfunctioning detectors.

	A	B	C	D	E	F
1	freeway	postmile	direction	type	lane	diagnosis
4247	101	4.84	S	ML	2	
4248	101	4.84	S	ML	3	
4249	101	4.84	S	ML	4	
4250	101	4.84	S	OR	1	card
4251	101	4.84	S	OR	2	
4252	101	4.84	S	OR	3	
4253	101	5.01	S	FR	1	
4254	101	5.01	S	FR	2	card
4255	101	5.01	S	ML	1	
4256	101	5.01	S	ML	2	
4257	101	5.01	S	ML	3	
4258	101	5.01	S	OR	1	
4259	101	5.01	S	OR	2	
4260	101	5.01	S	OR	3	
4261	101	5.09	N	ML	1	card
4262	101	5.09	N	ML	2	
4263	101	5.09	N	ML	3	
4264	101	5.09	N	ML	4	
4265	101	5.09	N	OR	1	
4266	101	5.09	N	OR	2	
4267	101	5.09	N	OR	3	card
4268	101	5.09	S	FR	1	comm
4269	101	5.09	S	ML	1	
4270	101	5.09	S	ML	2	
4271	101	5.09	S	ML	3	
4272	101	5.09	S	ML	4	
4273	101	5.17	N	ML	1	
4274	101	5.17	N	ML	2	
4275	101	5.17	N	ML	3	
4276	101	5.17	N	OR	1	open
4277	101	5.17	N	OR	2	card
4278	101	5.17	N	OR	3	open

Figure 2.36: Report of loops and error causes.

The PeMS report in Figure 2.36 also gives a hint as to why the loops aren't working. Sometimes, the cause is easy to diagnose, such as when there are no data coming from a detector. This means either the communications link is down between the loop and the TMC, or the controller is not working. On the other hand, when we do get data but they are all zeros, then the problem is not communications, but most likely in the electronics of the detector card. The PeMS diagnosis is summarized in Table 2.6. PeMS declares these

Error Type	*Description*	*Symptom*
communications	loss of communication between TMC and detector	no data from a detector station
detector card	detector electronics malfunction	all zero outputs for a particular lane
loop cut	loop wire is cut, creating an open circuit	high occupancy, zero flow
electrical connection	bad electrical connection due to moisture, bad splicing, etc	intermittent problems

Table 2.6: Physical causes of errors.

types of errors based on tests on the data. The last column in Figure 2.36 shows the type of error for each loop. If this field is empty, then the loop is good. These diagnoses enable the repair person to know what to expect before he goes on a field visit. For example, he may give the ones with communication errors to a communication specialist, and only visit those that have electronic errors. The intelligence provided in PeMS will greatly improve the efficiency of loop repair and maintenance. This is a vital part of an electronic surveillance system.

2.5 Automated traveler information

PeMS data and processing enable it to serve as an Advanced Traveler Information System (ATIS). An important goal of ITS is to inform the driver of the best routes and times of travel. While it is difficult to eliminate traffic delays, an ATIS can save time and reduce stress. PeMS contains a travel time prediction service accessible on the web. Using this service, the user enters the starting point, destination, and a desired departure or arrival time, and PeMS computes travel times on all possible routes and finds the best ones. Current measurement and historical relationships are used to predict travel times. This service not only indicates the good routes, but also greatly reduces the uncertainty in the delay. The detail of travel time prediction is given in Chapter 7.

Figure 2.37 shows the interface of this service. The user chooses starting and destination points from a list of freeway to freeway crossings. The application quickly determines the best route for this trip and displays the results on a Web page shown in

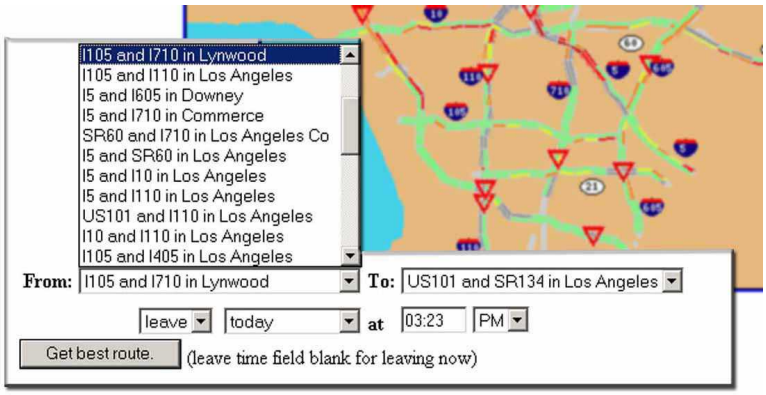


Figure 2.37: Route selection for travel time prediction.

Figure 2.38. This page provides a list of alternate routes and their travel times as well as

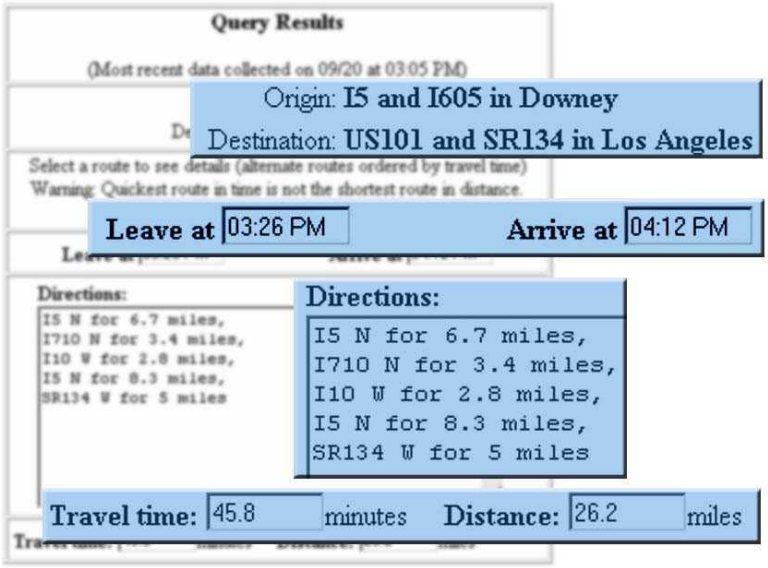


Figure 2.38: Result of query

the best route. Any departure time in the future can be specified. The prediction accuracy depends on the departure time, but is accurate up to one hour in advance.

performance of prediction A prototype of this service is also implemented in WAP [17], and similar services can be easily implemented as a voice service and or for in-car navigation devices.

Chapter 3

Custom Applications And Results

PeMS's built-in, Web-based applications are convenient for day-to-day use, but sometimes they are not enough for in-depth research. The most flexible way to access PeMS's data is through database queries. We obtained several important results about traffic behavior using custom experiments on historical data.

PeMS represents a new way of doing transportation research. Traditionally, researchers based their results on data from a handful of locations. But there is value in studying a large number of locations to have a statistical understanding of phenomena, since results from a single location may not be representative. Part of the problem is that there has been a lack of wide-covering traffic data. We used PeMS data to analyze traffic flow in many location and across long time periods in several studies.

In this chapter, we describe the results of three studies using data from PeMS. The first deals with freeway capacity at bottleneck locations. We find bottleneck locations and times in the Los Angeles region over one month, and analyze their capacity both under

free flow and bottleneck modes. We discover that in almost all locations, capacity was greater during free flow than during congestion. The second study measures travel time variability on one freeway route. Using both loop and incident data, we calculate the travel time statistics on this route, and estimate the real cost of travel in terms of scheduling time. The third study defines a methodology for measuring recurrent vs. non-recurrent delay, and applies it to PeMS data.

3.1 Freeway efficiency during free flow

Congestion describes the condition on the freeway when speed is low and density (in vehicles per mile) is high, and the traffic is "stop and go." California's urban freeways are usually congested during peak travel times. The observed flow rates during congestion is much less than the observed capacities at the same locations during free flow, which means that freeways operate inefficiently during times of most demand. Free flow describes the condition of high speeds, high flow rates, and moderate density. The transition from free flow to congested flow is often followed by a drop in the observed flow rate [15] [18] [19] [16].

The higher flows during free flow produce higher VMT, while the higher speeds mean less VHT is consumed. Since freeway efficiency is defined as the ratio of VMT to VHT, efficiency is maximized when free flow is maintained. Because congestion often results from merging traffic from on-ramps, ramp metering can be used to maintain free flow conditions on the mainline freeway sections. Studies have shown that ramp metering reduces or eliminates congestion, increases flow rate, and reduces delay and travel time

variability [20], [21].

This research investigates the gains in speed and flow rates when free flow is maintained.

3.1.1 Speed and flow rate are highest during free flow

The relationship between flow rate, density, and speed is described by several fundamental diagrams of traffic. Flow rate increases with density up to a critical density, then decreases if density rises further. This relationship is observed at many locations. Although PeMS measures occupancy, not density, occupancy is related to density and the average vehicle length by

$$\text{Density} = \frac{\text{Occupancy}}{\text{average vehicle length (ft)}} \times 5280 \text{ feet per mile.}$$

Figure 3.1 shows the observed relationship between 5-minute flow and occupancy samples at

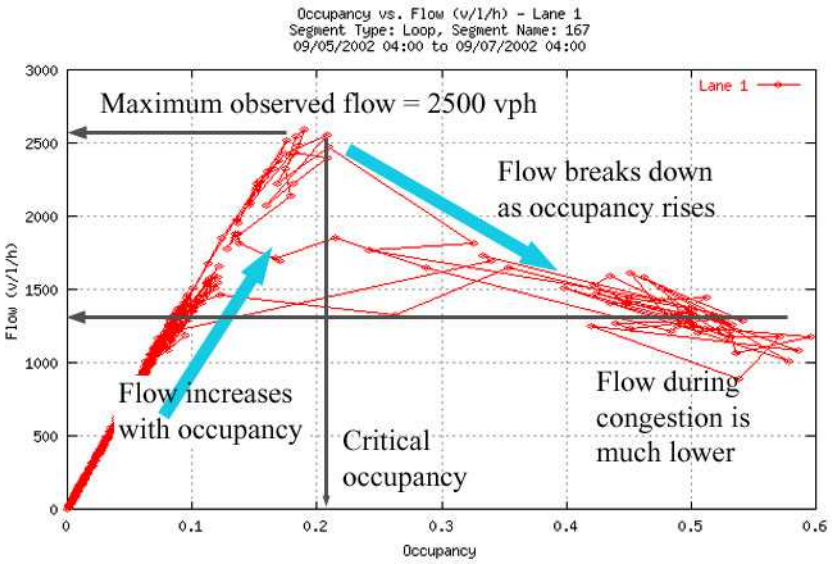


Figure 3.1: Fundamental diagram of traffic showing relationship between flow and occupancy.

a detector location on I-805N in San Diego, at postmile 22.4 in lane 1 between 9/5/2002 and 9/7/2002. The maximum observed flow rate at this location is more than 2500 vph, which was achieved when occupancy = 0.2. Flow rate drops markedly as occupancy increases further. During congestion, the flow rate is about 1300 vph.

The slope of Figure 3.1 is proportional to the speed. Figure 3.2 shows that during congestion, speed drops as well as flow. The x-axis shows the flow rate in vehicles per hour

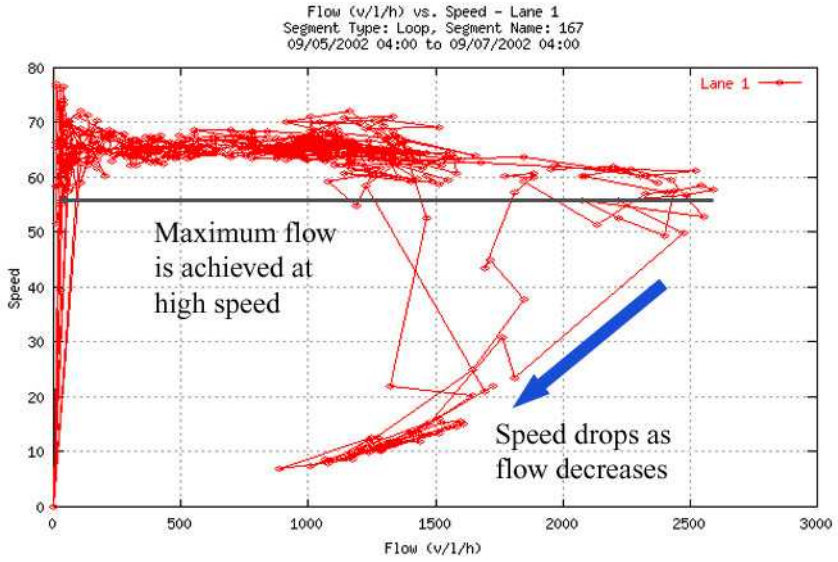


Figure 3.2: Speed versus flow.

and the y-axis shows the speed in miles per hour. At this location, during free flow, the flow rate is 2500 vph at a speed of 60 miles per hour; in congestion, the flow rate is between 1000 and 1500 vph and speed is between 10 and 20 mph. Since flow is proportional to VMT and speed is inversely proportional to VHT, the difference in efficiency between these operating points is

$$\frac{2500 \times 60}{1000 \times 20} = 7.5.$$

Therefore this location is 7.5 times as efficient during free flow as it is during congestion.

That the maximum flow rate is achieved during free flow is observed at other locations. Figure 3.3 shows the distribution of speed at locations in Los Angeles during periods of greatest observed flow. This plot is based on data from 2349 loops on 9/1/2000.

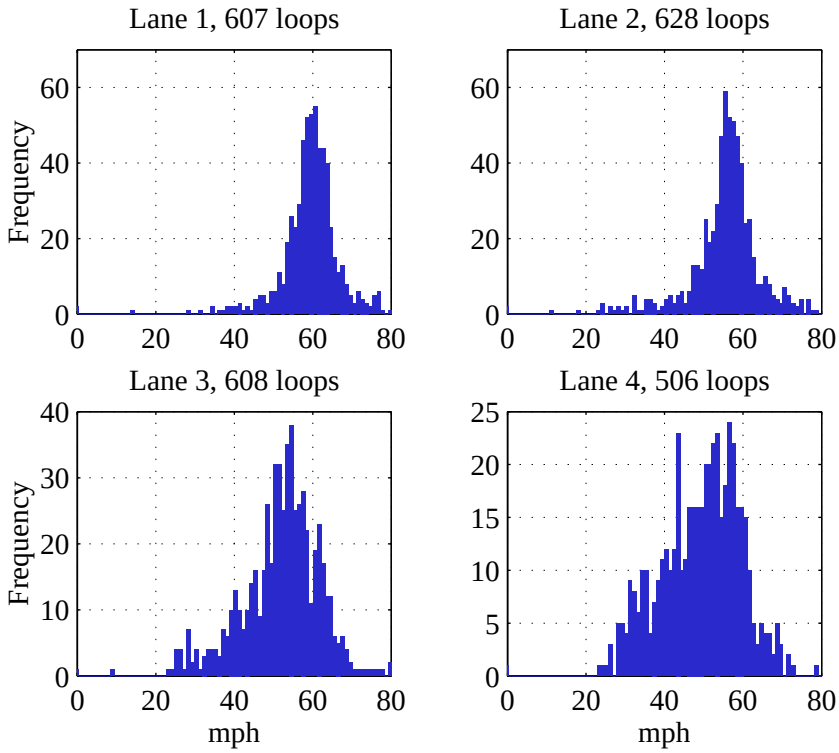


Figure 3.3: Speed distribution in lanes 1-4 in Los Angeles during periods of maximum observed flow.

Loops whose maximum flow rates are below 1500 vph are eliminated because they can probably accommodate more flow if there is more demand. Lanes 1 and 2, the two left lanes, have sharp peaks at 60 and 55 mph, the free flow speeds in those lanes. This shows that in most locations, maximum flow rate is achieved during free flow. Similar peaks are also visible in lanes 3 and 4, but they are not as distinct as in the left lanes This may be because there are more slow vehicles getting on and off the freeway in lanes 3 and 4.

In contrast to Figure 3.3, the speed distribution when traffic is at the highest observed occupancy on this day is shown in Figure 3.4. Speed at most locations are below 20 mph,

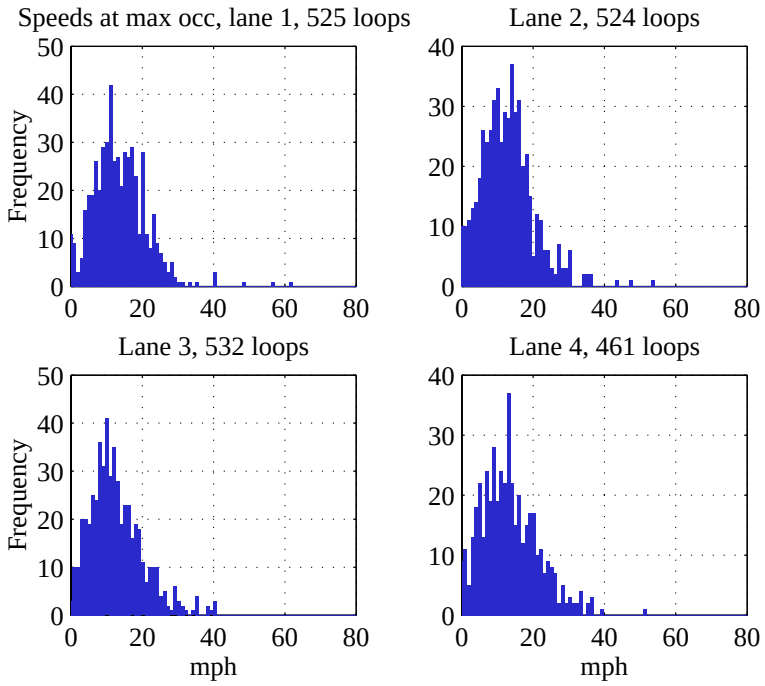


Figure 3.4: Speed distribution during highest occupancy.

compared to 60 mph in free flow. These speeds correspond to travel times three times higher than those in free flow.

The above evidence shows that free flow is more efficient than congested flow, therefore free flow should be maintained. This may be accomplished by ramp metering, which controls the rate of on-ramp merging traffic to regulate freeway density. Several studies performed by state DOTs show significantly improved efficiency from metering. The Minnesota Ramp Metering Study compared the performance before and after ramp metering was *removed* in the Twin Cities in 2000. It found that without ramp metering, flow rates

were reduced by 14%, and overall travel time became longer and twice as unpredictable [21]. Washington State DOT performed an evaluation on ramp metering effects on SR-520 East between I-5 and the entrance of the Evergreen Point Bridge, a distance of 2 miles. Ramp metering eliminated stop-and-go congestion and increased flow rates by 14% during peak hours [20]. These studies show that ramp metering reduces congestion and increases throughput in the real world.

3.1.2 Bottleneck capacities

Congestion usually starts from the same locations on freeways. Figure 3.5 shows that on I-805N in San Diego, congestion forms at the same place almost everyday. This

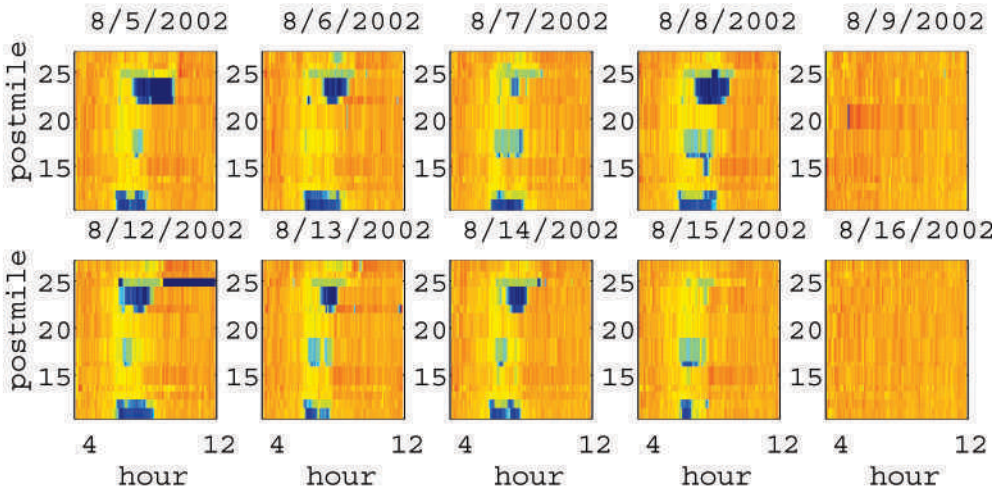


Figure 3.5: Contours of speed during two weeks on I-805N in San Diego, between 4:00 AM and 12:00 PM.

figure shows speeds on 10 weekdays between 8/5/2002 and 8/16/2002, where the dark regions represent low speeds. On eight of the days, congestion forms at postmile 25 and postmile 12. Postmile 25 is near a freeway entrance where traffic merges from SR-52. See

Figure 2.20. The chronic congestion in Figure 3.5 may be caused by bottlenecks. Once congestion starts, the region of congestion propagates upstream and reduces the efficiency of a large section of the freeway. It can even spill onto other freeways [20]. Therefore, it's important to prevent congestion from forming at these bottleneck locations. The following study finds that bottlenecks flows are also more efficient during free flow, therefore free flow should be maintained to increase throughput and prevent congestion.

Unlike in a queuing system, capacity at a given freeway location depends on the traffic pattern. Hall [15] observed that at one location he studied, the maximum observed flow rate is higher when there was no queue than the maximum observed when there was a queue. This suggests that there the capacity under free flow is higher than in a bottleneck at the same location. We call the flow rate when there is a queue the *discharge rate*, this is the rate at which vehicles at the head of the queue are released. We call the difference between the two capacities R, the *free flow gain*. The numerical value of R is important because of its implications on ramp metering. If R is large, then it may be worthwhile to meter the flow to achieve the higher flow rate. Hall measured R of 5% - 6%. Banks [16] performed a similar study at four locations, and found free flow gains of -1.2% to 3.2%. He concluded that this difference is not big enough for ramp metering because of the inefficiencies in any practical metering system.

While the studies of Hall and Banks examined the two-capacity phenomenon, they are based on data from a small number of locations. These locations were chosen by the researchers and may not be representative. Furthermore, no mathematical or statistical characterization of a bottleneck is given, nor is such a characterization to be found in other

transportation literature. The choice of study locations could only be based on subjective ideas on how a bottleneck should behave. In this section, we present a mathematical characterization of a bottleneck. Using this characterization, which we implemented in an algorithm, we found 23 persistent bottlenecks on nine freeways in Los Angeles. The free flow gains at these bottlenecks vary between -3% and 16%, with a median of 4.5%. Except for one location, free flow capacities are higher than bottleneck discharge capacities.

3.1.3 Bottleneck's role in capacity

In the Highway Capacity Manual (HCM), The capacity of a freeway section is defined as the maximum flow rate that can be sustained for 15 minutes [22]. To be conservative, we use 25 minutes as the condition for "sustained". If demand exceeds capacity, congestion will form, and the section becomes an active bottleneck. We use the word "active" to stress that the same location can act like a bottleneck or not, depending on demand.

A bottlenecks can form where traffic from an on-ramp merges with the main line flow, as illustrated in Figure 3.6. Suppose the downstream section, labeled "D", has a

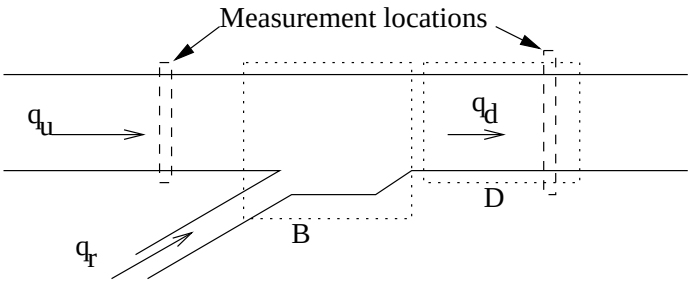


Figure 3.6: Bottleneck caused by an on-ramp.

capacity of y vehicles per hour. The demand flows are $q_r(t)$ and $q_u(t)$ at time t. When demand exceeds capacity, a queue forms in the area labeled "B", at the rate of $y - (q_r(t) + q_u(t))$ vehicles per hour. While this system is similar to a queuing system. real traffic behaves slightly differently. When there is a queue in "B", the flow in "D" is called *discharge* flow; when there is no queue, the flow is called *free flow.* It has been observed that the sustained discharge flow rate is often lower than the maximum measured sustained free flow rate [15] [19]. This phenomenon is often called "breakdown" in transportation literature [18], and it is one of the arguments for ramp metering. If we can avoid breakdown by restricting the flow into region "D", then we may be able to keep the flow rate at close to the free flow capacity [23].

To incorporate breakdown into our model, we replace the capacity y by two capacities y_f and y_b to represent the free flow and discharge capacities. The flow rate is $q_d(t) = q_u(t) + q_r(t)$ while $q_u(t) + q_r(t) \leq y_f$. When y_f is exceeded, congestion forms in "D" and the flow rate becomes $q_d(t) = y_b$, and the excess demand is stored in a queue that grows upstream.

This model still doesn't describe the real behavior of traffic because it is deterministic, whereas Hall observed that the actual capacities change from day to day [15], as did we. Therefore a stochastic description is needed. We replace y_b and y_f with random variables $Y_b(d)$ and $Y_f(d)$, where d is the index of day. They are the maximum achievable sustained flow rates on the dth day in discharge and free flow modes, respectively. Hall studied one location and showed that Y_b and Y_f have normal distributions, and that

$$\frac{\mathrm{E}(Y_f - Y_b)}{\mathrm{E}Y_b} \approx 5\% \tag{3.1}$$

Banks [16] performed a similar study on 4 locations, and found this ratio to be between -1.2% and 3.2%. Cassidy [19] observed an 8% gain.

We use PeMS data in the following study to verify these results and find a statistical description of freeway capacity. Let $Y_b(d,x), Y_f(d,x)$ be the observed capacities on day d and location x, since capacity depends on the freeway location. The free flow gain

$$R(d,x) = \frac{Y_f(d,x) - Y_b(d,x)}{Y_b(d,x)} \tag{3.2}$$

is a random variable for each d and x. We assume that $R(d,x)$ is a random variable i.i.d. in d. The studies in [16] and [15] demonstrate the need to understand how $R(d,x)$ varies over many locations x.

3.1.4 Estimate capacities

Bottleneck definition

By definition, the discharge capacity, $Y_b(d,x)$ is always achieved at an active bottleneck, therefore it can be measured if the bottleneck can be identified. Despite its conceptual importance, a mathematical characterization of a bottleneck is not found in literature. While Banks used a speed drop to identify the onset of congestion [16], Hall used occupancy to detect the formation of a queue upstream of the bottleneck[15]. We offer our own definition of a bottleneck, for these reasons. First, we need a way to find bottlenecks using an algorithm. Second, having a mathematical characterization permits comparison of results from different studies. Third, and most important, using a mechanical procedure guarantees that we are objective in our choice of study locations. We need to know that whatever we find out about bottlenecks is not a part of their definition.

Intuitively, when a bottleneck is active, the speed immediately downstream is high, while the upstream speed is low. This leads to our definition of a bottleneck. Let x_u, x_d be a pair of immediate upstream-downstream locations; let $V(d,t,x_u), V(d,t,x_d)$ be the average speed of all lanes at these locations at time t on day d. We use the notation of d,t to specify time, where $d = 1, 2, \ldots$ is the day and $t = 0, 1, \ldots, 287$ is the sequence of 5-minute samples of each day. We offer the following definition of a bottleneck.

Definition 1 *At time d,t, the pair of locations x_u, x_d include an active bottleneck if*

1. *There is a sustained speed drop between downstream and upstream, i.e.*

$$V(d,\tau,x_d) - V(d,\tau,x_u) > \delta, \ \forall 0 \leq \tau - t < \theta;$$

where $\delta \stackrel{\text{def}}{=} 15$ mph, $\theta \stackrel{\text{def}}{=} 5$ (25 minutes)

2. *Upstream is in congestion, i.e. $V(d,t,x_u) < v_{cong}$, where $v_{cong} \stackrel{\text{def}}{=} 50mph$.*

We implemented an algorithm to find bottlenecks that conform to the above definition. Bottlenecks can be identified visually from a contour plot such as the one in Figure 3.7. The dark regions represent low speeds, and traffic moves in the direction of increasing postmile. This figure shows several bottlenecks. For example, at mile 15, between 15:00 and 17:00, and between 17:30 and 20:30, there is an active bottleneck. Contour plots show both the duration and location of active bottlenecks.

The same contour plot is superimposed with the bottlenecks detected by our algorithm, which are marked with upside-down triangles. This algorithm acts on each sample time t to find all the active bottlenecks at that time. For example, the speeds at every location at 18:45 are shown in Figure 3.8, where the direction of travel is from left to right. This

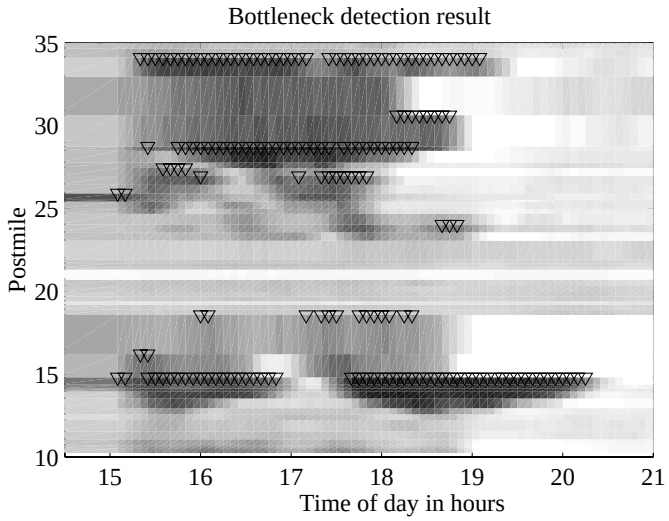


Figure 3.7: Speed contour for a given day on I-10 East in Los Angeles.

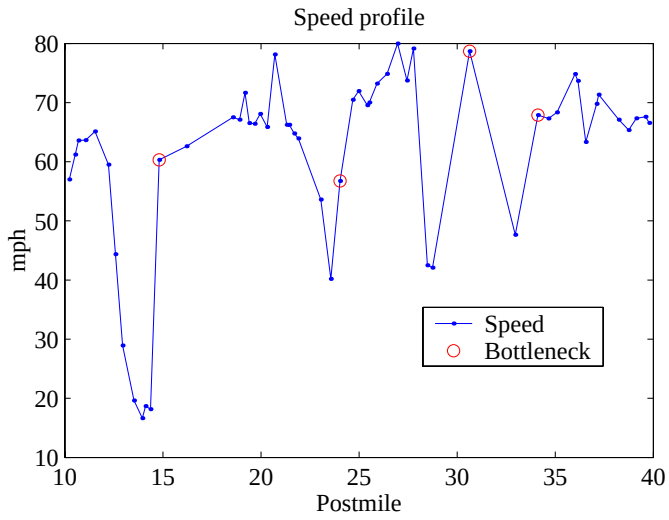


Figure 3.8: Speed profile, and bottleneck detection.

is a vertical "slice" of the contour plot at t =18:45 in Figure 3.7. The algorithm detected four active bottleneck locations for this time, at postmiles 15, 24, 30.5, and 34, marked by circles in Figure 3.8. We found the times and places of all the bottlenecks on this day, and

marked them on the speed contour in Figure 3.7 with triangles. The result of the detection algorithm agrees well with visual inspection.

We measure the discharge flow at the detected bottlenecks. However, not all bottleneck flows are discharge flows. The following section explains why bottlenecks formed by off-ramps and accidents do not have discharge flow, and must be excluded from the analysis.

Other types of bottlenecks

Figure 3.6 shows that a potential bottleneck at an on-ramp location. Bottlenecks can also form at off-ramp locations, but the flow rate downstream of an off-ramp bottleneck is not the discharge flow. Figure 3.9 shows an example of this, also called a diverging bottleneck. When the demand for the off ramp is greater than its capacity, congestion

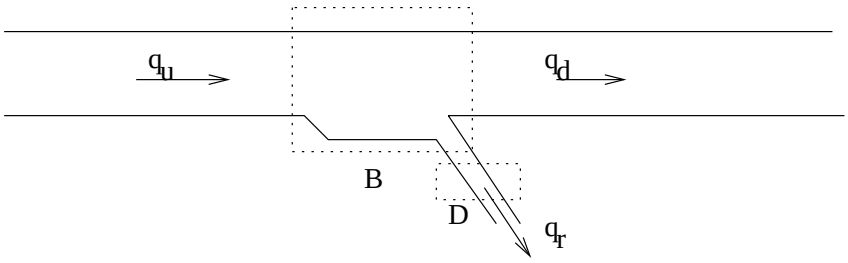


Figure 3.9: Potential bottleneck created by off-ramp.

forms in "B". But the flow rate q_d is not the discharge rate. Rather, q_r is the discharge rate of this bottleneck. This is not the situation we are studying.

To eliminate off ramp bottlenecks from the data, we add an extra condition to specify on-ramp bottlenecks only. Again let x_u, x_d be a pair of upstream-downstream locations. The modified algorithm declares (x_u, x_d) to be an active bottleneck only if the

flow downstream is greater than upstream. If (x_u, x_d) contains a diverging bottleneck, the flow downstream has to be less than upstream, because some vehicles exit at the off-ramp; if an on-ramp caused the bottleneck, then downstream flow is the sum of upstream and the on-ramp flow. Of course, we can also visually observe the geometry at each location to determine whether it is a diverging bottleneck, but it's much easier to implement the condition on flow as an algorithm.

Bottlenecks can also be caused by incidents. For example, an incident that blocks lanes at a location lowers the capacity can cause queueing upstream. However we are interested in the recurring reduction in flow because of congestion when all lanes are available. We remove incidents from our analysis, according to the CHP incident database.

Free flow capacity

By removing accidents and off-ramp bottlenecks, we can be fairly certain that the discharge capacity $Y_b(d, x)$ is achieved downstream of an active bottleneck. However, we never know if the free flow capacity $Y_f(d, x)$ is achieved. By definition, $Y_f(d, x) \geq Q(d, t, x)\ \forall (d, t)$ in free flow, where $Q(d, t, x)$ is the measured 25-minute flow rate on day d, at time t at location x. But we cannot measure Y_f directly, only its lower bound $\hat{Y}_f$, where

$$\begin{aligned} \hat{Y}_f(d, x) &\stackrel{\text{def}}{=} \max_{\{(d,t) \text{ is in free flow}\}} \{Q(d, t, x)\} \\ &\leq Y_f(d, x) \end{aligned} \tag{3.3}$$

with equality only when there is enough demand upstream. We use (3.3) to estimate Y_f, but realize that it's a lower bound.

3.1.5 Method and results

Method

We use data from 20 weekdays in August 2002 on nine freeways in Los Angeles, listed in Table 3.1. They are chosen because they lead to downtown LA. On each freeway,

Freeway	*Direction*	*Length(mi)*	*Locations*
I-5	N	78	89
I-10	W	47	110
SR-60	W	30	37
US-101	S	49	68
I-110	N	23	35
I-210	W	25	39
I-405	N	44	81
I-605	N	24	43
I-710	N	13	21
Total		333	523

Table 3.1: Capacity study locations.

there are detectors at locations $x_1, \ldots, x_k, \ldots$. We label the days sequentially with $d = 1, 2, \ldots$. On each day d and at each location x, we examine 5-minute average speed $V(d, t, x)$ and average flow rate $Q(d, t, x)$ between 6:00 am and 12:00 pm. We label time using $t = 0, 1, \ldots$, where each index is a 5-minute sample period, with $t = 0$ corresponding to 6:00 am.

Using the data and the definition of an active bottleneck, we find times and locations where bottleneck or free flow conditions existed. Let $B(d, t, x)$ be the indicator that the location x is the *downstream* of an active bottleneck at (d, t), using Definition 1. Let

$F(d,t,x)$ be the indicator of free flow, where

$$F(d,t,x) = 1 \text{ iff } \begin{cases} B(d,t,x) = 0 & \text{and} \\ \\ V(d,t,x) \geq v_f \end{cases} \tag{3.4}$$

where $v_f \stackrel{\text{def}}{=} 50\text{mph}$.

The free flow capacity was defined loosely in (3.3). It is the maximum sustained flow rate when the location is in free flow. We define "sustained" as at least 25 minutes, or 5 consecutive 5-minute sample periods. We estimate $\hat{Y}_f(d,t)$ using

$$\hat{Y}_f(d,x) = \max_{\{t:\, F(d,t+i,x)=1,\, 0\leq i\leq \theta-1\}} \left\{ \frac{1}{\theta} \sum_{i=0}^{\theta-1} Q(d,t+i,x) \right\} \tag{3.5}$$

where $\theta = 5$. The discharge capacity estimate is

$$\hat{Y}_b(d,x) = \max_{\{t:\, B(d,t+i,x)=1,\, 0\leq i\leq \theta-1\}} \left\{ \frac{1}{\theta} \sum_{i=0}^{\theta-1} Q(d,t+i,x) \right\}. \tag{3.6}$$

In these equations, $Q(d,t,x)$ is the flow rate of location x, which is the sum of flows on all lanes at that location. Unfortunately, some PeMS locations are missing data for some lanes, but we need the total flow rate across all lanes. Though we developed a way of imputing missing data (see Chapter 5), it was not available for this study. Instead, we approximate the total flow by multiplying the average flow of the available lanes by the total number of lanes at each location.

Results

Using this method above, we found bottleneck locations and when they were active. There are 24 locations that have sustained bottlenecks on at least 5 out of the 20 days studied. For each of these locations x_i, we estimate the average free flow gain:

$$\hat{R}(d,x_i) \quad = \quad \frac{\hat{Y}_f(d,x_i) - \hat{Y}_b(d,x_i)}{\hat{Y}_b(d,x_i)}$$

$$\bar{R}(x_i) \quad = \quad \frac{1}{n_d} \sum_{d=1}^{n_d} \hat{R}(d, x_i), \; n_d = 20 \text{ days.} \tag{3.7}$$

Figure 3.10 shows that $\bar{R}(x_i)$ has a range of between -3% and 15%, except for the location

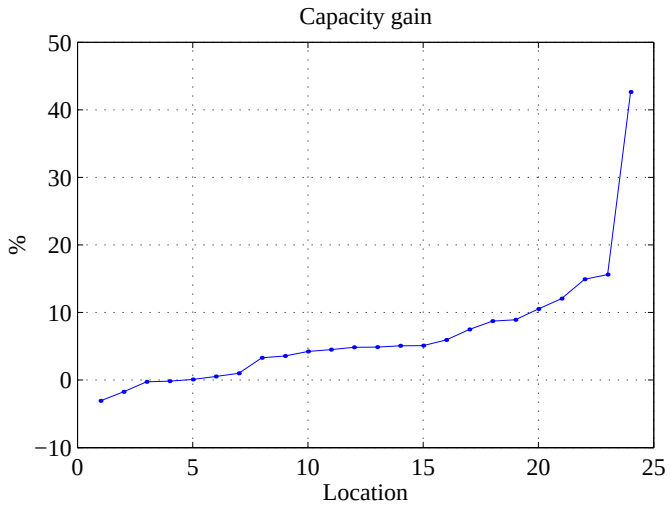


Figure 3.10: Free flow gain of 24 locations (sorted).

where $\bar{R}(x) = 43\%$. A closer look at the data revealed that this location is a diverging bottleneck. It was not automatically detected because the number of lanes dropped from 4 to 3 between upstream and downstream, causing the algorithm to overestimate the downstream flow. This location is discarded.

In 22 out of 23 remaining bottleneck locations in Figure 3.10, $\bar{R}(x_i)$ is positive, which means the flow rate is higher during free flow than during congestion. The median free flow gain of the 23 locations is 4.5%, which is consistent with the studies in [16] and [15]. However, about a third have free flow gains above 5%. Therefore in these places, the gain is significant and should be exploited to maximize total flow rate.

An individual location

We examine some locations more closely to check our results. There are several sources of error in the study. For example, is our detection algorithm in fact identifying bottlenecks? Are we estimating the capacities correctly? The data we used, while having wide coverage in time and space, contain many bad samples. Are our results tainted by bad data? To answer these questions, we look at several locations in detail to verify that we are measuring the desired quantities.

Since the last point in Figure 3.10 is an error, we take the second to last point, with the second highest gain at about 15%, and examine this location in detail. This location is on Interstate 210 West, at postmile 30.139. We look at one day of speed and flow from Thursday, 8/1/2002, at both upstream and downstream locations. There are 4 lanes at this location, but because of missing data, there is data only lanes 2 and 3 in the upstream, and lanes 1 and 2 in the downstream.

Figure 3.11 shows speeds upstream and downstream of this bottleneck on this day, where the upstream is the lower of the two curves, and the squares mark the active bottleneck times found by our algorithm. Between 7:00 and 7:30, speed upstream drops from 50 mph to 20 mph, while the downstream speed mostly remained above 50 mph. The algorithm determined there was an active bottleneck at these times. Figure 3.12 shows the flows up and downstream in the same time period, where the top curve is the downstream flow, and bottleneck times are marked by squares. It shows that the flow rate during bottleneck times is about 1970 vphpl on average. On the other hand, the maximum sustained free flow rate was 2359 vphpl, between 6:00 and 7:00 The free flow gain for this day is therefore

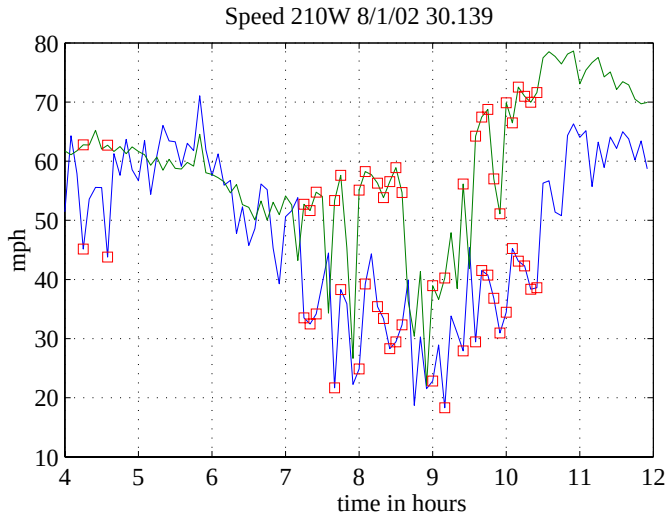


Figure 3.11: Speed upstream and downstream of bottleneck. Downstream is showing higher speeds.

$(2359 - 1970)/1970 \times 100 = 20\%$.

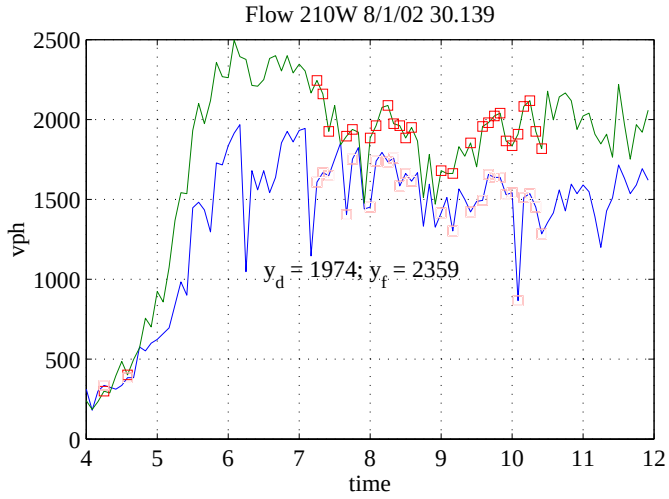


Figure 3.12: Average lane flow rate upstream and downstream of bottleneck. Higher flow is downstream.

The algorithm properly detects these drops in speed and flow rate. But we need to verify (1) was there an accident that choked off the flow at 7:30, (2) is this a diverging bottleneck, and (3) what is the effect of the missing data? In response to (1), a query into

the incident database revealed no incidents at this location and these times. Regarding (2), this location does not appear to be a diverging bottleneck because the flow downstream is greater than the flow upstream. Second, there are no major freeways to diverge onto at this location. Third, this is the morning rush hour and traffic is generally moving toward the east toward downtown.

We address the issue of missing data. Only data from lanes 1 and 2 of downstream, and lanes 2 and 3 of upstream are available. The observed free flow gain is based only on the left two lanes in the downstream. Banks found that a 10% free flow gain in the left most lane translated into a smaller gain of 3% in the total flow. Therefore, we expect the overall gain across all four lanes to be somewhat less than the observed 20% on this day.

3.1.6 Conclusion

We find that freeways operate at maximum efficiency during free flow. However, during periods of peak demand, freeways are usually congested and operate at a much lower efficiency. Efficiency is defined by the ratio of VMT to VHT. During free flow, speeds are about 60 mph, or three times higher than during congestion, so VHT is reduced by two thirds. Flow rates are also higher during free flow than in congestion in most locations, producing higher VMT. Congestion usually starts at the same bottleneck locations on a freeway. Ramp metering should be implemented at these locations to prevent congestion formation on the mainlines. Maintaining free flow at bottlenecks also increases their throughput. In Los Angeles, the free flow gain is observed to be between -3% and 15%. The combined gain in flow rate and speed makes a strong case for ramp metering.

3.2 Travel time as a measure of service quality

Travel time and its variability are important to drivers. Incidents and recurrent congestion increase travel time and also make it more unpredictable, costing the driver in expected and unexpected delays. Therefore, travel time statistics indicate the quality of service received by drivers.

Direct travel time measurements are not often available because they require probe vehicles. PeMS computes travel times from loop detector speed estimates using an algorithm described in Section 3.2.2. We present the travel time behavior on I-5N in Los Angeles, and show that travel time quantifies the service quality of a freeway route. We show that the quantile travel times such as the 90th percentile is a meaningful way to combine the travel time mean and variance.

Travel time can be used to gauge the benefits of Intelligent Transportation Systems (ITS). While it is desirable to reduce overall travel time, it is also beneficial to improve the predictability of delay. We show that the knowledge of incidents improves the predictability of travel time.

3.2.1 Measures of LOS

The Highway Capacity Manual (HCM) defines six levels of service based on density of vehicles per mile per lane. The purpose of this measure is for "operational analysis, design, and planning" [22]. LOS is a local measure, and is used to analyze the operation of specific locations, such as a weaving section [24]. It is also used to design freeways [25]. For example, a freeway designer must consider the demand volume, the grade of the road,

and merging traffic, and design the geometry, number and width of lanes, shoulders, etc to provide acceptable service quality most of the time.

VMT, VHT, and delay measure system performance in terms of output, input, and congestion cost. Travel time is also a measure of performance. While the total travel time of all vehicles for a given period is equal to the VHT, travel time variability is an extra cost that isn't measured by VHT or delay. Studies have found that people place a cost not only on the average travel time, but also its variance. According to a survey [26], drivers placed a cost of \$2.6 - \$8 per hour of total delay, and \$10-\$15 per hour of the standard deviation of delay. The variance of travel time manifests itself in the scheduling cost, which is the extra time one must budget for a trip because of its uncertainty.

The cost of travel is more than simply the average travel time because often, people have to arrive on time. For example, a truck driver who must reach the factory before closing time needs to budget extra time for unforeseen congestion delays. Therefore the maximum travel time, or, more realistically, the 90th percentile [27] is the real cost. We will show that the 90th percentile travel time meaningfully combines the mean and variance of travel time, both in terms of interpretability and as a way of measuring dollar cost.

3.2.2 Calculating travel time from loop speeds

Given speed at time t and location x on a route, we can estimate the travel times by computing vehicle trajectories [28]. Let $\tilde{V}(i,j)$ be the measured speed at discrete times t_i and locations x_j, and $V(t,x)$ be the actual speeds for all times and locations, where $\tilde{V}(i,j) = V(t_i, x_j)$ for all i, j. Given a departure time s_0, and a starting location y_0 on the

freeway, a vehicle that obeys $V(t,x)$ has the trajectory $y(t)$, where

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = V(t, y(t)),\ y(s_0) = y_0. \tag{3.8}$$

To compute the travel time of a vehicle with a given departure time, we need to estimate its trajectory which obeys (3.8). PeMS obtains speed estimates every five minutes from detector stations that are about a half mile apart. From this discrete set of measurements, we first approximate a continuous speed surface in the time-distance plane, and then estimate a trajectory by "walking" on this surface iteratively. Let $\hat{V}(t,x)$ be approximation to the actual speeds $V(t,x)$. $\hat{V}(t,x)$ is found by interpolating between the four measurements closest to (t,x) in space and time. Define a distance function $d(t,x,s,y)$ as

$$d(t,x,s,y) \stackrel{\text{def}}{=} \sqrt{(t-s)^2 + \left(\frac{x-y}{v_o}\right)^2}, \tag{3.9}$$

where $v_o = 45$ miles per hour. v_o weights time and distance such that the influence of a measurement Δ hours ago at the same location is equal to that of a measurement at the same time but at Δv_o miles away. $\hat{V}(s,y)$ is interpolated from the four nearest measurement points to (s,y) using the distance function. Let $(i_-,j_-),(i_-,j_+),(i_+,j_-),(i_+,j_+)$ be the 4 nearest measurement points, such that

$$i_+ - i_- = 1,\ t_{i_-} \le s \le t_{i_+},\ j_+ - j_- = 1,\ x_{j_-} \le y \le x_{j_+}. \tag{3.10}$$

Label these points $(i_1,j_1),\ldots,(i_4,j_4)$. Given any (s,y), we estimate the speed using

$$\hat{V}(s,y) \stackrel{\text{def}}{=} \begin{cases} \frac{\sum_{k=1}^4 \tilde{V}_k \frac{1}{d_k}}{\sum_{k=1}^4 \frac{1}{d_k}} & \text{if } d_k > 0 \text{ for all } k \\ \tilde{V}(i_k,j_k) & \text{if } d_k = 0 \text{ for some } k \end{cases} \tag{3.11}$$

where $d_k \equiv d(s,y,t_{i_k},x_{j_k})$ and $\tilde{V}_k \equiv \tilde{V}(i_k,j_k)$. $\hat{V}$ allows us to estimate a discrete trajectory $(s_0,y_0),(s_1,y_1),\ldots$, where each successive point is computed from the previous point using

the speed at that point, using

$$y_{m+1} = \hat{V}(s_m, y_m)(s_{m+1} - s_m) + y_m. \tag{3.12}$$

The travel time of a trip between locations y_0 and y_{dest}, departing at time s_0, is found by counting the time it takes for the discrete trajectory to reach the end point.

Figure 3.13 illustrates the travel time estimation process by showing the measured speed field and computed trajectories for several departure times. Dark regions represent low speeds, and trajectories are shown as white lines going from bottom to the top, in the direction of travel. The slope of the trajectories is small where speeds are low.

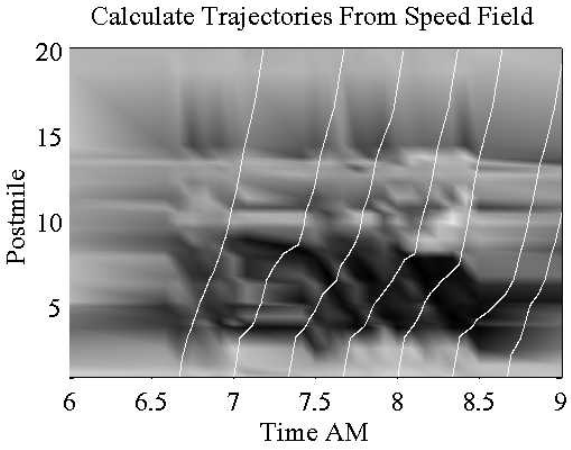


Figure 3.13: Trajectories, computed from speed field.

The PeMS travel time estimates are accurate when compared against direct measurement from Caltrans TACH probe vehicle runs in August 2002 in Los Angeles. Fifteen runs were made on 8/10/2000 on one route, and another 11 on 8/16/2000 on another route. The comparison is shown in Figures 3.14 and 3.15. On 8/10/2000, the probe travel time has two peaks at 4:30 pm and 5:30 pm and our estimates reproduced the second peak well,

but missed the first. On 8/16/2002, PeMS estimates agreed well with measurement. On 8/10/2002, PeMS did not receive data from several detectors on this route, so speed for those locations were interpolated from the other detectors. This is likely the cause of the missed congestion. These comparisons show that PeMS can reproduce TACH measurements very well – as long as the data are available.

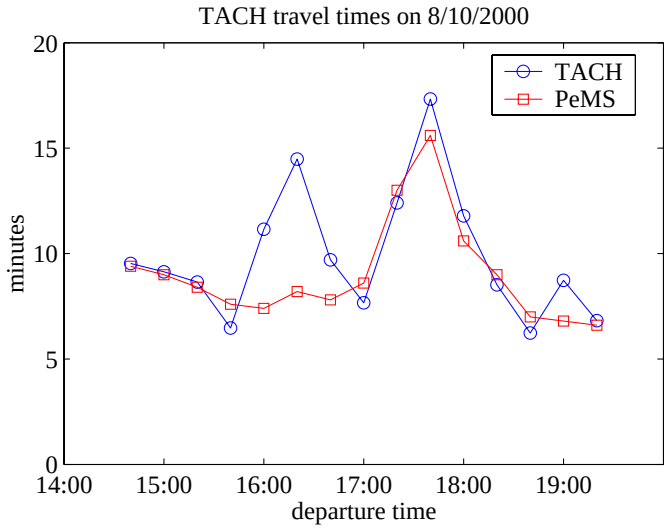


Figure 3.14: TACH runs on 8/10/2002

Travel times on I-5 corridor

We now present the travel time statistics on a section of I-5 North in Los Angeles County. The raw data are in the form of speed versus time and location. They are provided by PeMS. The corridor in this study is 20 miles long, between postmiles 0 and 20. We computed the travel time on this corridor on 65 weekdays between 3/1/2002 and 6/1/2002, between 5 am and 10 pm each day. Their mean, 10th and 90th percentile are plotted in Figure 3.16 for each departure time of day. As expected, travel time is higher during

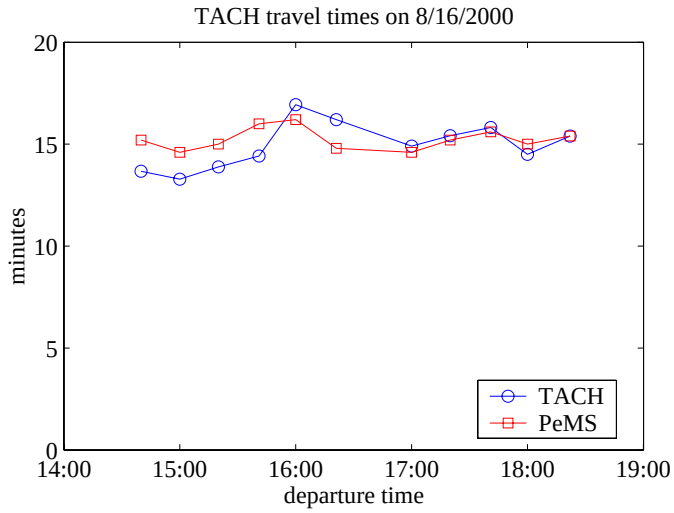


Figure 3.15: TACH runs on 8/16/2002

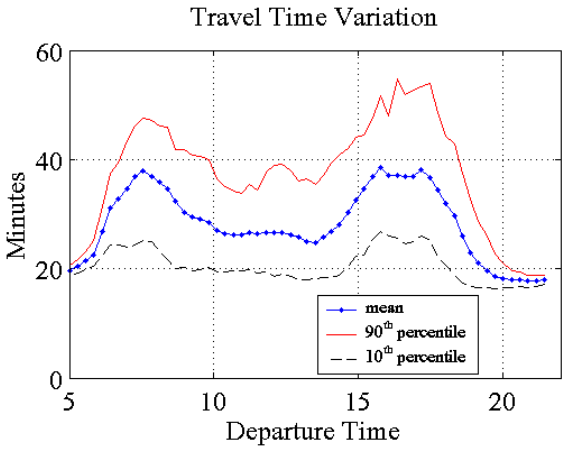


Figure 3.16: Mean, 10th and 90th percentile travel times.

morning and evening peak hours. However, travel time variability is high throughout the day with a range of 20 minutes between 10th and 90th percentiles. These statistics are useful to drivers. For example, to arrive at the destination 6 pm with 90% probability, one must leave 55 minutes earlier; to arrive at noon, one only needs to budget 40 minutes.

3.2.3 Quantifying the cost of travel

The standard deviation is commonly used to describe the variability of a random variable. The standard deviations of travel times for each departure time of day are plotted in Figure 3.17; Figure 3.18 shows the relationship between σ and μ. The standard deviation

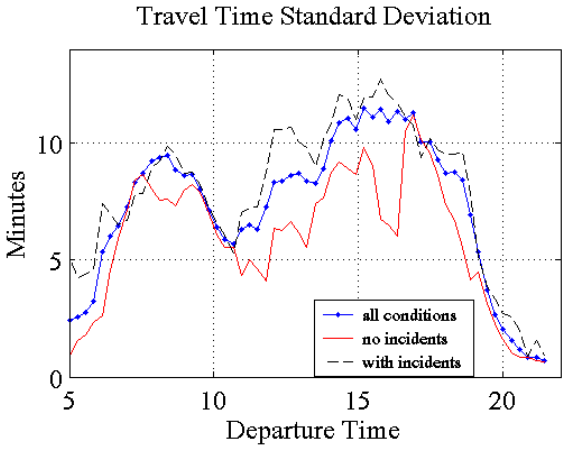


Figure 3.17: Travel time standard deviation.

increases with mean travel time, as we expect. In the literature, the standard deviation, along with mean travel time, are frequently used to calculate the cost of driving in a corridor [27][21], by

$$c = r_{avg}\mu + r_{std}\sigma \tag{3.13}$$

where

c	=	total cost
μ	=	average travel time
σ	=	standard deviation of travel time
r_{avg}	=	cost per unit time of average travel time
r_{std}	=	cost per unit time of standard deviation of travel time

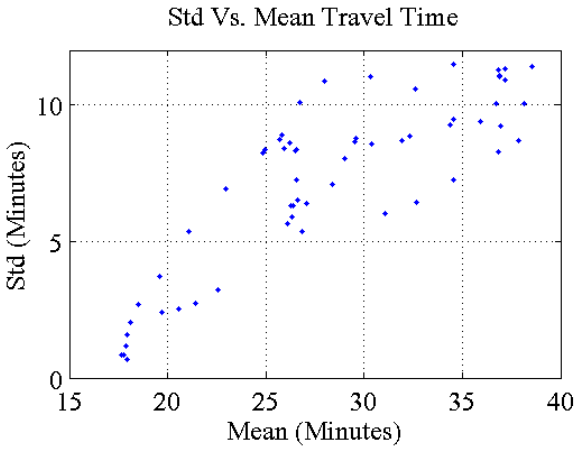


Figure 3.18: Standard deviation versus mean travel time.

Equation (3.13) combines the costs of average travel time and travel time variability. The cost per unit time r_{avg} and r_{std} are obtained from surveys. A more interpretable formulation takes the 90th percentile travel time as the cost of travel. This is because the 90th percentile is the amount one must budget to have a 90% chance of arriving on time. So the cost of travel for a trip with uncertain travel time is

$$\text{cost} = (\text{unit cost of time}) \times (\text{90th percentile travel time}).$$

The use of the 90th percentile, rather than 80 or 95, is somewhat arbitrary. Ideally, the full distribution is more informative. But we chose 90% because 1) a single number is easier to convey than a distribution; 2) a much higher percentile requires more data; 3) a much lower percentile doesn't represent the real cost. The 90th percentile travel time meaningfully combines travel time mean and variance into one number.

Impact of incidents

Incident increase the mean and variance of travel time. We measure the effect of incidents on travel time and show that real time incident information improves the predictability of travel time. This is a benefit of Intelligent Transportation Systems (ITS).

We use incident information from the California Highway Patrol (CHP) website [1]. An incident can be any number of events, such a vehicle collision, a stalled vehicle, debris on road, etc. Each incident has a start and an end time, a classification, and a location. In this study, we don't differentiate among different types of incidents. In the future, it will be useful to identify the impact of different types of incidents.

The incidents are correlated with trips, which are the trajectories calculated from measured speeds, like the trajectories in Figure 3.13. We computed trajectories for trips departing every 17 minutes from the origin, on each of 65 weekdays. A trip is said to be an *incident trip* if at any time during the trip, there is an incident anywhere on the study section. For each departure time of day, there are 65 trips, one for each day. Each trip was classified as an incident trip or a non-incident trip. The fractions of incident trips are plotted by departure time in Figure 3.19. This plot shows, for example, that a trip departing at 10 am has a 60% chance of witnessing an accident.

As expected, both the standard deviation and the median of travel time are larger when there are incidents, as shown in Figure 3.17 and Figure 3.20. These figures show that incidents have a measurable effect on travel time - the cost of incidents is about 5 minutes per vehicle when there is an incident. The combined effect of increases in average and variance of travel times is captured in Figure 3.21, which shows the 90th percentile travel

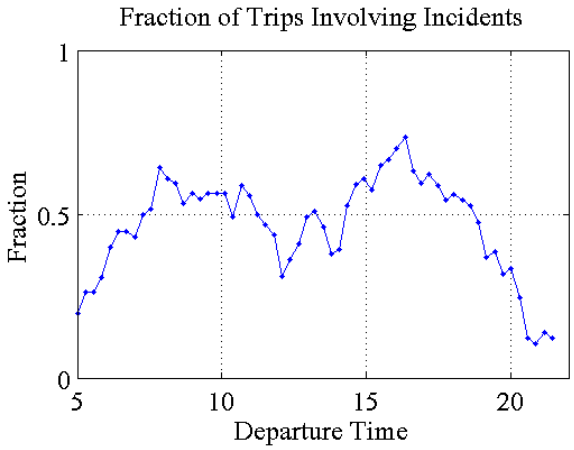


Figure 3.19: Fraction of trips with incidents.

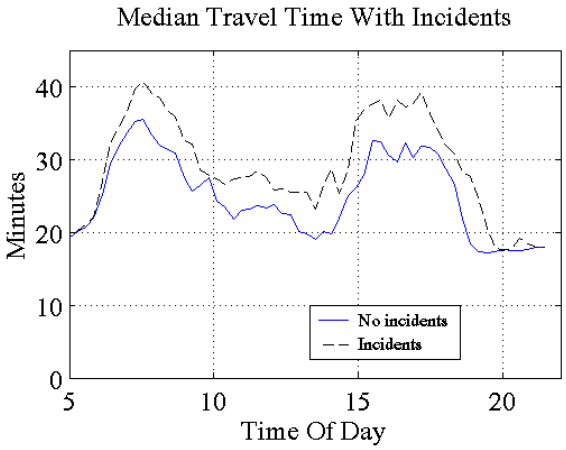


Figure 3.20: Median travel times under incident and non incident conditions.

time under incident and non incident conditions. Figure 3.21 has valuable information for the driver. For example, if one wants to arrive at an appointment at 3 pm, and there are no reported incidents, it's enough to budget 35 minutes. On the other hand, when incident information is not available, the 90th percentile time is 45 minutes, according to Figure

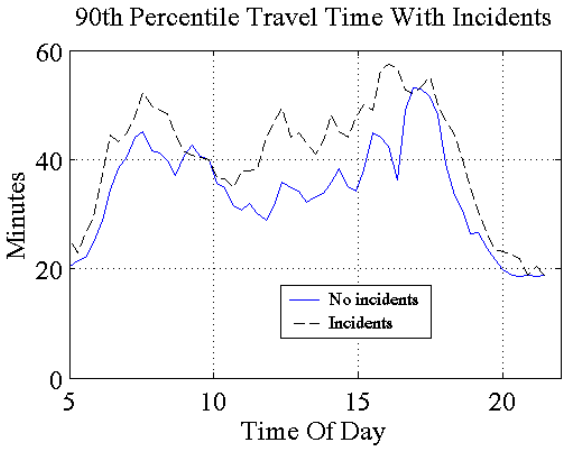


Figure 3.21: Ninetieth percentile travel times under incident and non incident conditions.

3.16. Therefore, knowledge that there is no incident provides a saving of 10 minutes in this case.

The example above demonstrates how ITS can improve the quality of travel by making it more predictable. An Advanced Traveler Information System (ATIS) that alerts drivers of incidents in real time can use the relationship between incidents and travel time to make a more accurate prediction of travel time. Similarly, knowing current speeds on the corridor also helps in predicting travel time, as demonstrated by van Zwet in [29]. His results show that travel time on a 40-mile corridor can be predicted with an root-mean-squared(RMS) error of 12 minutes when the current speeds are known, compared to a error of up to 30 minutes when no information is available. These findings clearly illustrate the benefit of ITS, and these benefits can be measured in travel time reliability.

3.2.4 Travel time and traditional LOS

The HCM defines the Level of Service (LOS) of a homogeneous road segment. The LOS is given on a scale from A to F based on vehicle density, defined in Table 3.2. LOS is used in the design and operation of freeway sections. An engineer designs freeway sections based on projected demand patterns, freeway geometry, and grades for a target LOS. The actual LOS achieved on a freeway is a measure of its performance.

LOS	Density (veh/mi/lane)
A	$K \leq 11$
B	$11 < K \leq 18$
C	$18 < K \leq 26$
D	$26 < K \leq 35$
E	$35 < K \leq 45$
F	$K > 45$

Table 3.2: Definition of LOS in HCM 2000.

We computed LOS from single loop occupancy measurements. The average length L, the occupancy O, and the density K of a sample period are related by

$$K = \frac{O}{L},$$

where L is in miles per vehicle, and K in vehicles per mile. L is estimated with the g-factor algorithm, see Chapter 6. LOS is related to travel time-we expect the travel time to be longer and more variable during congestion. Figure 3.22 confirms this intuition. It shows the mean and standard deviation of travel time under each level of service in Los Angeles. This plot uses data recorded on 4/24/2002 between 6 and 8 am. Under LOS A-E, the travel time is low and consistent; under LOS F, the travel time is 3 times longer than free

flow, and the standard deviation is also much greater. This suggests that LOS of A-E are acceptable levels of service, while LOS of F is unacceptable.

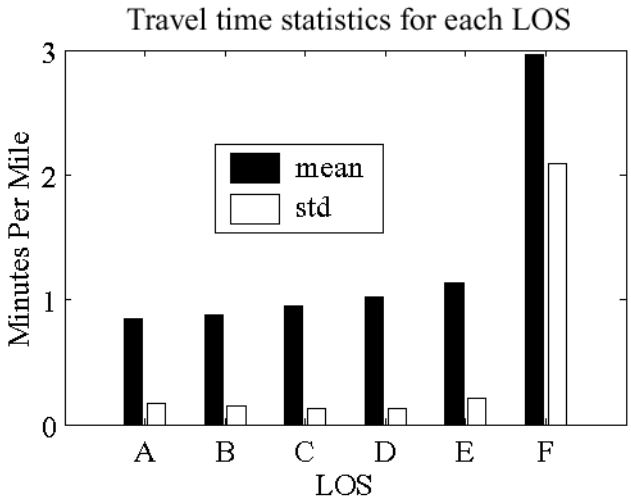


Figure 3.22: Travel time variability for each LOS in HCM.

Figure 3.23 shows the fraction of the corridor where the LOS is F, for each time of day. To produce this plot, we first calculate the LOS for each detector location x on the corridor (there are about 30), at each time t, for each day d (there are 22 days). Let $L(d,t,x)$ be this quantity. For each sample time we calculated the percent of the freeway that had LOS F:

$$P_F(d,t) = \frac{1}{n_x} \sum_{a \le x \le b} 1(L(d,t,x) = F), \tag{3.14}$$

where a, b are the boundaries of the corridor and n_x is the number of detector locations. Next, we compute the average of $P_F(d,t)$ over the days,

$$R_F(t) = \frac{1}{n_d} \sum_{d=1}^{n} P_F(d,t), \tag{3.15}$$

where $n_d = 22$ is the number of days. Figure 3.23 shows that the amount of congestion is greatest during peak travel times in the morning and evening, a familiar pattern that is also exhibited by the travel time. For example, on average, 50% of the length of a trip at 8:00 am is congested; at noon, only 25% is congested. For the traveler, however, the effect of increased congestion is best understood in terms of time. Figure 3.16 shows that a driver will save 10 minutes by traveling at noon instead of 8 am.

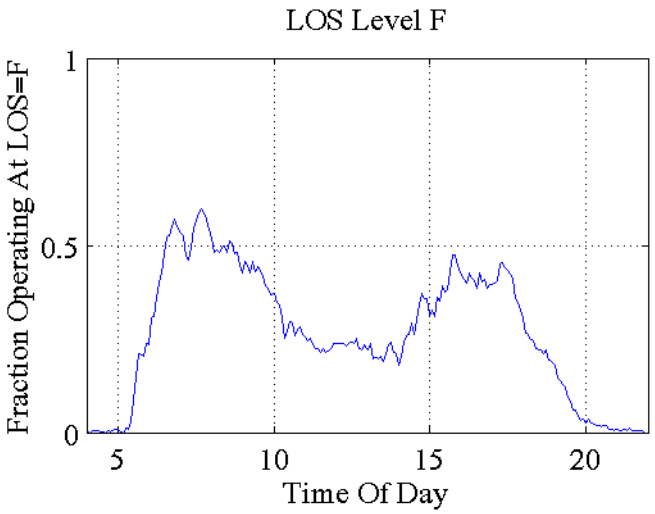


Figure 3.23: Fraction of corridor at LOS of F at various times of day.

3.2.5 Conclusion

We present the travel time behavior of a busy Los Angeles corridor using loop data. Travel time quantifies the level of service experienced by the driver and is more interpretable than the LOS defined by the HCM. LOS is given in discrete levels and are not easily translated into dollar cost. In addition, LOS is a local measure and does not easily

describe the experience of a driver on a trip consisting of many freeway segments.

To directly measure travel time requires probe vehicles. However, PeMS produce accurate estimates of travel times for freeway routes using loop speed estimates. We present travel time statistics for a 20-mile corridor. Because of the high variability of travel time, an average of 10 minutes must be added to the average travel time to ensure 90% chance of on-time arrival for most departure times.

PeMS incident data were used to estimate the effect of incidents on travel time. We found that on the study corridor, advance knowledge of an incident can reduce the scheduling time by about 10 minutes. This is a direct measure of the benefit of ITS.

3.3 Recurrent/non-recurrent delay

Incidents create delays and uncertainty in travel. The previous section quantified the impact of incidents on travel time and its variability. While travel time is the key quantity for an individual driver, the DOT is also interested in reducing the overall delay, measured in vehicle-hours. For example, this can be achieved by clearing incidents faster and focusing efforts on the most delay-causing incidents. The total delay of a region can be divided into recurrent and non-recurrent. Recurrent delay arises from fluctuations in demand, the manner in which the freeway is operated, as well as the physical layout of the freeway. Non-recurrent delay depends on the nature of the incident: an accident is likely to cause more delay than a vehicle stopped on the shoulder of the highway.

Currently, there are several approaches for defining and measuring congestion delay. For example, Caltrans defines the total delay in a freeway section as the additional

vehicle-hours traveled driving below a reference speed (e.g., 35 mph). Recurrent delay is measured using probe vehicles to record travel times during incident-free periods. Non-recurrent congestion is usually assumed to be equal to the recurrent congestion. Other congestion-related performance measures include travel rate, percent facility segments with demand higher than capacity, or threshold speeds. In general, however, there is a lack of consistent definition and measurement of the congestion and its components using real-world data

We present a methodology to identify and measure recurrent and non-recurrent delay using loop data. Using this method, we found a statistical description of these types of delays on a freeway corridor in Los Angeles.

3.3.1 Method

This section presents the statistical model and empirical procedures used to estimate recurrent and non-recurrent congestion on freeways, based on PeMS data.

The basic quantity of interest is the random delay due to congestion in a highway section s over a time period t. Denote this random delay over a section-duration pair (s,t) by $D(s,t)$. $D(s,t)$ is measured in the PeMS system as the excess vehicle-hours traveled below a reference speed. More precisely:

$$D(s,t) = \sum_{\{\sigma \in s, \tau \in t\}} \max\{\text{VHT}(\sigma, tau) - \frac{\text{VMT}(\sigma,\tau)}{V_r}, 0\}. \tag{3.16}$$

Here, σ indexes a PeMS segment (i.e. a section of highway half-way between two consecutive detector stations), τ indexes a 5-minute PeMS average quantity, $\{\sigma \in s, \tau \in t\}$ is the set of 5-minute segment-intervals belonging to the (s,t) pair, $\text{VMT}(\sigma,\tau)$ is the vehicle-miles

traveled and VHT(σ, τ) is the vehicle-hours traveled over the segment-interval , and V_r is the reference speed-either 35 mph or 60 mph.

Formula (3.16) says that $D(s,t)$ is the excess vehicle-hours spent by vehicles over the section-duration pair (s,t) traveling at a speed below V_r mph. Observe the effect of temporal and spatial granularity in the PeMS data: aggregating over larger segments or averaging over longer time intervals (say, 15-minute) will lead to lower measured delays.

By accepting that the delay is a random quantity, we are also accepting that a single sample measurement of the delay – as is commonly done by measuring the delay experienced by a single probe vehicle run – does not provide a meaningful estimate of this delay. For one example segment (considered below) over the 33 days during February-April, 2002, for which there were no incidents during the morning peak period 06:00-10:00 am, the delay ranged from a minimum of 0 veh-hrs (VH) to a maximum of 1,098 VH, with a mean of 322 VH and a standard deviation of 255 VH.

Because this delay is random, our objective is to obtain a statistical characterization of this delay. Such a characterization may include statistical mean, variance, quartiles, and probability distributions.

We also want to separate this delay into the recurrent delay – the delay that occurs in the absence of incidents; and non-recurrent delay-the additional delay caused by incidents. Moreover, we may wish to allocate the non-recurrent delay to individual factors. Because of the limitations imposed by the CHP incident data in our empirical study, we only consider two factors: accidents and non-accident incidents.

We disentangle recurrent from non-recurrent delay and estimate the impact of

different kinds of incidents with the help of a statistical model:

$$\mathrm{P}[D(s,t) = d] = \sum_{I} \mathrm{P}[D(s,t) = d|I]\mathrm{P}[I]. \tag{3.17}$$

In this equation, $\mathrm{P}[D(s,t) = d]$ is the probability that the random delay $D(s,t)$ equals d; I denotes the type of incident; $\mathrm{P}[D(s,t) = d|I]$ is the probability that $D(s,t) = d$, conditioned on the occurrence of an incident of type I; and $\mathrm{P}(I)$ is the probability of occurrence of such an incident.

In the empirical study, we distinguish between $I = 0$, I =acc or accident, and I =non or a 'non-accident' incident.

The empirical study has two limitations. First, in studying the delay over a particular (s,t) pair, we include only those incidents that occur within the (s,t) pair. This can cause two kinds of errors. The first limitation might be called the 'boundary effect'. Suppose an incident occurs within a section-period pair (s,t). In our study, we estimate the impact of this incident in terms of $D(s,t)$. But the incident's impact could extend to a section s' downstream of s or to a period t' after t. (In both cases, the impact would be counted as 'recurrent' congestion.) However, our empirical study does not attribute this delay to the incident that occurred in (s,t). Thus we must be careful in choosing the size of the sections and the durations to be large enough so that this boundary effect is relatively small. In our empirical study this 'boundary effect' is minimized because s is taken to be long sections (several miles) of freeway and t is a long duration-the peak travel time.

The second limitation is due to coverage. We limit ourselves to incidents reported in the CHP/CAD database, because these are the data we can collect on-line. We know that this does not include all incidents. However, the accidents in the CAD appear to

Incident type	In-lane	Shoulder	Total	CHP total
Accident	16	42	58	54
Breakdown	67	532	599	6
Debris	27	17	44	67
Total	110	591	701	127

Table 3.3: FSP and CHP incident records for I-210, per month

match well the accidents reported by the freeway service patrols (FSP). Table 3.3 shows a comparison between the two data sources for the I-210 test section in Los Angeles. Only the FSP records show whether an incident is blocking a lane or on the shoulder. The most underreported incident is in the 'breakdowns' category. Also, from the analysis of incident data on I-10 and I-880, for which we have detailed incident data from observers in probe vehicles, we find that CHP/CAD data includes only 15-20 percent of all incidents, but it does include virtually all accidents and all the delay-causing incidents [30]. Other causes of non-recurrent congestion include lane closures, events, and inclement weather. In principle, these could be included in (3.17), simply by considering them as new kinds of incident. In the application of the methodology, we ignore these causes because of lack of data.

We can use (3.17) to decompose the expected value of the total delay, $\mathrm{E}[D(s,t)]$, into recurrent and non-recurrent delay:

$$\begin{aligned}
\mathrm{E}[D(s,t)] &= \sum_d d \times \mathrm{P}[D(s,t)=d] \\
&= \sum_I \sum_d d \times \mathrm{P}[D(s,t)=d|I]\mathrm{P}[I|s,t] \\
&= \sum_I \mathrm{E}[D(s,t)|I]\mathrm{P}[I|s,t] \\
&= \mathrm{E}[D(s,t)|I=0]\mathrm{P}[I=0|s,t] + \sum_{I\neq 0}\{\mathrm{E}[D(s,t)|I] - \mathrm{E}[D(s,t)|I=0]\}\mathrm{P}[I|s,t] \\
&= \text{Recurrent congestion} + \text{Nonrecurrent congestion} \qquad (3.18)
\end{aligned}$$

In the second to last equation, we have used the assumption that

$$\mathrm{P}[I=0|s,t] = 1 - \sum_{I\neq 0} \mathrm{P}[I|s,t]. \tag{3.19}$$

For our empirical analysis, this means that if there is no CHP website report of an incident during the segment-duration pair (s,t), then there is in fact no incident.

Thus, the basic relations that we will estimate are:

$$\begin{aligned} \text{Total congestion} &= \text{Recurrent congestion} + \text{Nonrecurrent congestion} \\ \text{Recurrent congestion} &= \mathrm{E}[D(s,t)|I=0] \qquad (3.20) \\ \text{Nonrecurrent congestion} &= \sum_{I\neq 0}\{\mathrm{E}[D(s,t)|I] - \mathrm{E}[D(s,t)|I=0]\}\mathrm{P}[I|s,t] \qquad (3.21) \end{aligned}$$

In addition to these statistical averages, we also wish to estimate the distributions

$$\mathrm{P}[D(s,t)=d|I] \tag{3.22}$$

As mentioned, in our empirical study, we only distinguish between non-incidents ($I=0$), non-accident incidents (I=non) and accidents (I=acc) and so the relation for non-recurrent congestion simplifies to

$$\begin{aligned} & \text{Nonrecurrent congestion} \qquad (3.23) \\ = \; & \{\mathrm{E}[D(s,t)|I=\text{acc}] - \mathrm{E}[D(s,t)|I=0]\}\mathrm{P}[I=\text{acc}] \\ & +\{\mathrm{E}[D(s,t)|I=\text{non}] - \mathrm{E}[D(s,t)|I=0]\}\mathrm{P}[I=\text{non}] \qquad (3.24) \\ = \; & \text{Congestion from accidents} + \text{Congestion from non-acc} \end{aligned}$$

Note in both (3.21) and (3.24), to evaluate non-recurrent congestion we have to deduct $\mathrm{E}[D(s,t)|I=0]$ because, by definition, non-recurrent congestion is the excess over recurrent

congestion caused by incidents. Equations (3.21), (3.22), and (3.24) form the basis of our empirical study.

3.3.2 Application of the methodology

This section presents the application of the methodology to two real-life freeway corridors. We explain the procedures we use in the empirical estimates of the quantities in (3.21) - (3.24) , and the additional assumption underlying these procedures.

I-210: an 11-mile section of freeway 210 in Los Angeles. The study area is between postmiles 32 and 43. Congestion delays were calculated for the AM peak period 6:00 to 10:00 AM for the period February to April 2002 (60 weekdays). Data on traffic conditions and incidents are from PeMS. The study section experiences heavy recurrent congestion in the WB direction in the AM peak.

I-880: This is a 6-mile freeway section located in the city of Hayward, Alameda County. Data on traffic volumes and incidents were provided by the I-880 FSP database [31]. Two datasets were used. The 'before' data set includes information for 20 weekdays for the AM and pm peak periods. The 'after' data set includes data for 24 weekdays for the AM and pm peak periods. 'Before' and 'after' refer to periods before and after initiation of Freeway Service Patrol service.

The statistical assumption is that of stationarity and independence. More precisely, we fix a section-duration pair (s, t) in which s denotes a particular section (e.g. I-210W between postmiles 32 and 43 in LA) and t stands for a fixed weekday period such as the AM peak, 06:00-10:00. Suppose we have measurements of congestion delay and incidents for N weekdays, $t_1, \ldots, t_N$. We assume that these N samples are independent and

identically distributed. With this assumption, we can use empirical averages and frequency counts to estimate the statistical averages and probability distributions in (3.21) - (3.24).

We partition the N samples into three classes: N_0 is the set of samples n for which CHP reports no incidents; N_{acc} is the set of samples for which CHP reports at least one accident; and N_{non} is the set of samples for which CHP reports at least one incident but no accident. We ignore the distinction between the occurrence of one incident and two or more incidents, because there are very few cases of the latter in the CHP website during a single AM peak duration. The distributions (3.22) are estimated by the frequencies:

$$\begin{aligned} \hat{\mathrm{P}}[D(s,t) \in \mathrm{Bin}(d)|i] &= \frac{\text{Number of samples in } N_I \text{ with delay in Bin}(d)}{\text{Number of samples in } N_I}, \\ I &= 0, \mathrm{acc}, \mathrm{non}. \end{aligned} \tag{3.25}$$

Here Bin(d) stands for a delay 'bin'.

The conditional means are estimated by

$$\hat{\mathrm{E}}[D(s,t)|I] = \frac{\sum_{\{n \in N_1\}} D(s,t)}{\text{Number of samples in } N_1}. \tag{3.26}$$

Above $\hat{\mathrm{P}}$ and $\hat{\mathrm{E}}$ are our estimates and equations (3.25) and (3.26) summarize how these estimates are calculated from the data samples $D(s,t)$, $n = 1, \ldots, N$. Figure 3.24 shows the three distributions, $\mathrm{P}[D(s,t)|I = 0]$, $\mathrm{P}[D(s,t)|I = \mathrm{non}]$, $\mathrm{P}[D(s,t)|I = \mathrm{acc}]$ for the 11 mile study section of I-210W (from pm 32 to 43), during the 06:00-10:00 AM peak period.

Table 3.4 provides some descriptive statistics on congestion delay for the I-210 site. In the table, the first column is the type of I, the second column is $\mathrm{P}(I)$, the third column is the estimate of the mean congestion delay conditioned on the incident type, σ is the standard deviation of the samples, Error is the standard error of the estimated mean, Max D is the maximum value of the delay, and Count is the number of samples.

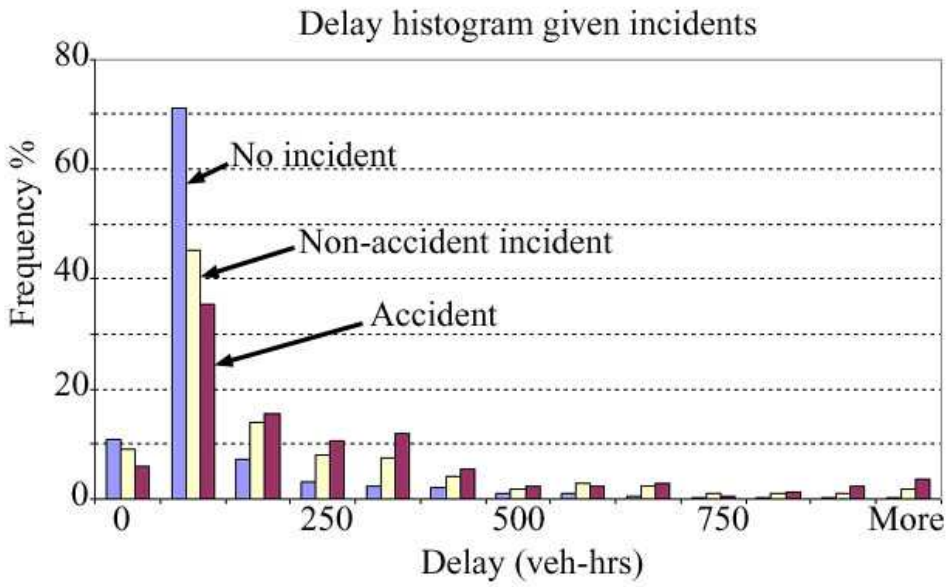


Figure 3.24: Conditional distribution of delay under incident and non-incident conditions, on I-210W, AM peak.

I	P(I)	E$\{D\|I\}$	σ	Error	Max D	Count
Total	1.00	368.75	290.67	18.53	1457.75	246
$I = 0$	0.66	322.00	255.00	19.97	1098.50	163
$I =$ inc	0.34	460.56	384.50	42.20	1457.75	83
$I =$ non	0.15	410.58	304.67	40.09	1271.00	37
$I =$ acc	0.19	500.75	352.75	52.01	1457.75	46

Table 3.4: Summary statistics using reference speed of 60 mph.

Several things are worth noticing. First, the estimate of (3.22) for 60 mph reference speed is

$$\begin{aligned} \text{Total congestion} &= \text{Recurrent congestion} + \text{Nonrecurrent congestion} \\ 368.75 &= 322 + 46.75 \end{aligned}$$

hence nonrecurrent congestion accounts for 46.75/368.75 or 13% of total congestion along the study corridor.

The nonrecurrent congestion breakdown is

$$\text{Nonrecurrent congestion} = \text{Congestion from } (I = \text{acc}) + \text{Congestion from } (I = \text{non})$$

$$46.75 = 13.25 + 33.50,$$

hence accidents account for 33.5/46.75 or 72% of nonrecurrent congestion.

Second, as is clear from Figure 1, as well as from the large standard deviation, the probability distribution of congestion delay has a large 'tail'. Consequently, measures of congestion must account for this variation. Giving a single number to summarize congestion is very misleading.

Table 3.5 shows the delay descriptive statistics for a reference speed of 35 mph, instead of 60 mph. This alters the quantitative conclusions above in two ways. The estimate of delay in each row of the table obviously goes down. Furthermore, the recurrent delay estimate ($I = 0$) will decline by a greater percentage, so that the percentage contribution of nonrecurrent congestion to total congestion will increase. From Table 3.5 we can see that the reference speed of 35 mph, non-recurrent congestion accounts for 36.58/214.41 or 17 percent of total congestion (vs. 12 percent for the 60 mph reference).

I	P(I)	E$\{D\|I\}$	σ	Error	Max D	Count
Total	1.00	214.42	196.50	12.53	1104.25	246
$I = 0$	0.66	177.83	166.42	13.03	806.17	163
$I = \text{inc}$	0.34	286.19	234.20	25.71	1104.25	83
$I = \text{non}$	0.15	251.92	205.17	33.73	842.83	34
$I = \text{acc}$	0.19	313.75	246.58	36.36	1104.25	46

Table 3.5: Summary statistics using reference speed of 35 mph.

Tables 3.6 and 3.7 show the results from the application on the I-880 test site for

both the before and after data sets. The results show that the percent of non-recurring congestion is about 28 to 30% of the total congestion. The same results were obtained when minor incidents (shoulder breakdowns lasting less than 10 minutes on the average) are excluded from the database. This indicates that the incident normally reported in the CHP/CAD account for most of the congestion delay.

I	P(I)	E$\{D\|I\}$	σ	Error	Max D	Count
Total	1.00	40.45	46.04	2.71	286.30	288
$I = 0$	0.22	28.30	28.07	3.54	167.30	63
$I =$ inc	0.78	43.85	49.44	3.30	286.30	225
$I =$ non	0.63	35.91	45.17	3.35	286.30	182
$I =$ acc	0.15	77.47	53.05	8.09	207.61	43

Table 3.6: Summary statistics of I-880, before FSP.

I	P(I)	E$\{D\|I\}$	σ	Error	Max D	Count
Total	1.00	47.94	49.86	2.91	244.31	293
$I = 0$	0.29	30.86	31.77	3.43	123.91	86
$I =$ inc	0.71	49.91	49.14	3.42	244.31	207
$I =$ non	0.57	43.82	47.49	3.69	240.10	166
$I =$ acc	0.14	74.51	48.57	7.59	244.31	41

Table 3.7: Summary statistics of I-880, after FSP.

3.3.3 Conclusion

Using the methodology defined in this section, we measured the amount of recurrent and nonrecurrent congestion on California freeways. Because of the variability in traffic, typical measurement techniques cannot fully capture the true range of conditions. Accurate measurement of delays is important and can be used in before-after studies to

evaluate the effectiveness of management strategies. For example, the results based on the before-and-after data from the FSP project suggest that it did not significantly reduce congestion.

We found nonrecurrent congestion to account for 13 to 30 percent of total delay, a smaller proportion than the 50% figure that is commonly accepted. This ratio, and the absolute amount of non-recurrent delay has an important impact on policies regarding freeway service patrol. Of course, this ratio may vary depending on the freeway and its demand. We are interested in urban, chronically congested freeways.

This study, along with the studies on capacity and travel time presented earlier in this chapter illustrates the need for statistical understanding of traffic congestion. We found congestion to be highly variable across locations and times when measured in flow rate, travel time, and delay. Therefore, automated surveillance is required to capture the amount of data needed for true analysis. We presented ways to statistically analyze these phenomena. The PeMS database allows traffic managers and researchers to perform these in-depth analysis on historical data.

Chapter 4

PeMS Components

In this chapter, we describe the three main functional units of PeMS – data collection, data processing, and data access. Traffic and other data are collected from various sources electronically and in real time. Raw data are processed and fused to calculate performance measures and other useful quantities. Users access PeMS applications primarily on the World Wide Web.

4.1 Data collection

The main source of PeMS data is loop detectors. They measure vehicle counts and freeway occupancy at thousands of freeway locations, across all lanes, and on on-ramps and off-ramps. Loop detectors use a wire coil buried under the freeway to detect vehicle presence, similar to a metal detector. The coil contains a few turns of a wire and have a diameter of about six feet, and is buried about one inch under the road surface. Figure 4.1 shows a picture of loops on a freeway in Los Angeles.

Figure 4.1: Loops on a freeway.

The important quantities to measure are speed, vehicle count, and freeway occupancy. There are many technologies used to measure traffic data, including radar, lasers, and magnetometers. PeMS is not restricted to any kind of data collection technology and can use data from many sources. PeMS uses loop data because loop detectors are the most abundant type of traffic sensors in California.

Loop detectors measure the change in inductance when a vehicle is present, which lowers the inductance of the loop [32]. The change in inductance is sensed by the detector card, an electronic device located in a nearby roadside cabinet and electrically connected to the loop. Figure 4.2 illustrates the detection process. A vehicle moves over the loop and causes its inductance to drop. When the inductance crosses a threshold, the detector logic switches from the "0" state to the "1" state. When the vehicle leaves the loop, the inductance rises to normal levels and crosses the threshold again. The passing of the vehicle

thus produces a pulse of detection, shown in the lower part of Figure 4.2. The duration of the pulse is the time it takes for the vehicle to move past the loop.

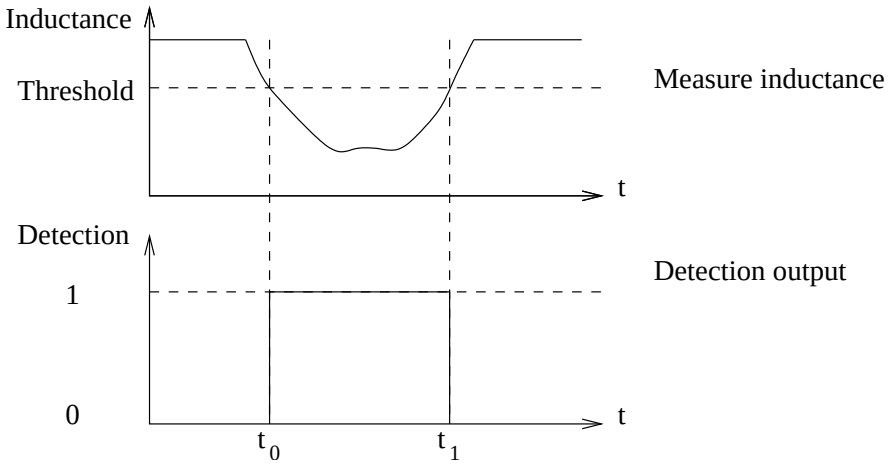


Figure 4.2: Detection using change in inductance in the wire loop.

Since each vehicle produces a pulse at the detector, as shown in Figure 4.2, a single loop detector measures vehicle flow by counting the pulses. It also measures the occupancy, which is the fraction of each sample period during which the detector output is "1". The sample period is typically 20 or 30 seconds. Occupancy is a measure of how much of the roadway is covered by vehicles. A related quantity is the freeway density, measured in vehicles per mile. It's important to note that the duration of each pulse depends on the vehicle length, speed, inductance signature, and the sensitivity of the detector card. Because of the dependence on detector card settings and loop electrical properties, the same vehicle with the same speed can produce different pulse durations at different locations. Therefore, for the same physical condition, different loops may register different measured occupancy if they have different sensitivities.

Speed is measured directly by double loop detectors, which use two loops close to

each other, as illustrated in Figure 4.3. The pair of loops are separated by a known distance

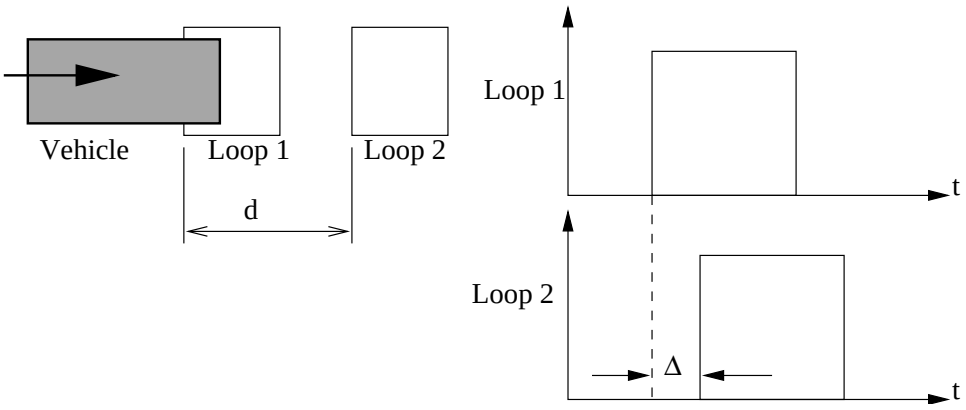


Figure 4.3: Double loop detection of speed.

d, which is on the order of 12 feet. A moving vehicle triggers the two loops one after the other. The time difference Δ between the two rising edges from the detection output of the two loops is used to compute the vehicle speed:

$$\text{speed} = \frac{d}{\Delta}.$$

Most of California's loop detectors are not double loops but single loops, which cannot measure speed directly. Speed can be estimated from single loops if the average vehicle lengths are known. This is a difficult problem because vehicle lengths at different locations and at different times are highly variable. PeMS uses an algorithm to estimate vehicle lengths and speeds. This algorithm is presented in Chapter 6.

Typically, there is a loop detector station at every ramp location. A loop is installed in each of the on-ramp, off-ramp, mainline, and HOV lanes. Loops are connected to a cabinet on the side of the road. See Figure 4.4. Inside the cabinet, each loop is connected to a detector card. The detector cards generate the detection pulses to be read

Figure 4.4: Cabinet location next to freeway.

by a computer called a controller. The controller gathers the vehicle count, occupancy, and speed data from the detectors every 30 seconds.

At the Caltrans Traffic Management Center (TMC), a computer called the Front End Processor (FEP) polls all the loop controllers every 30 seconds, and receives from them measurements of the most recent 30-second sample period. The FEP communicates with the controllers over phone lines, fiber links, or wireless connections. This process is illustrated in Figure 4.5. The FEP is on the local area network at the TMC, and feeds data to other Caltrans machines. It sends data to PeMS via the Caltrans Wide Area Network (WAN). PeMS runs on a Sun 450 computer called Transacct.eecs.berkeley.edu. It has four processors, two GB of RAM, and several terabytes of disk space. Disks are added as capacity is needed. The developmental system is housed in Cory Hall, EECS department, UC Berkeley; a production system has been installed at Caltrans District 4 District Office

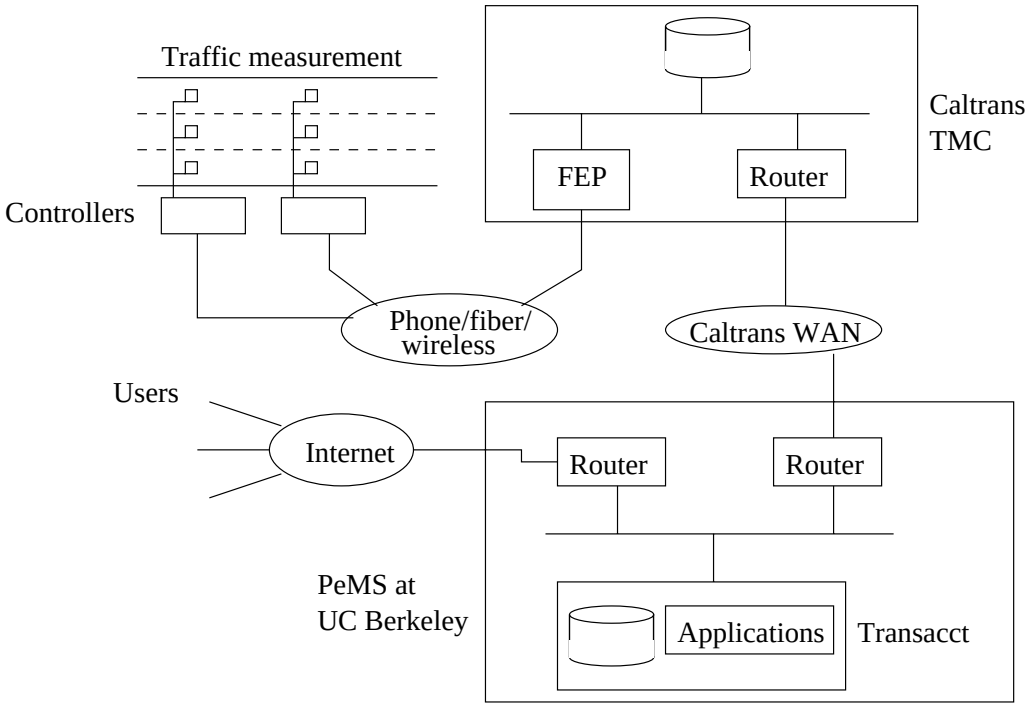


Figure 4.5: Data path between controllers to PeMS.

in Oakland. Transacct runs an Oracle database and the algorithms for data processing and serving users. Most users access PeMS via the internet using a Web browser.

PeMS collects loop data from six of 12 Caltrans Districts, covering all major California metropolitan areas. See Table 4.1. As of November 2002, there were 7522 loop

District	Name	Locations
3	Sacramento	199
4	SF Bay Area	1241
7	Los Angeles	3293
8	San Bernardino	470
11	San Diego	285
12	Orange County	2034
Total		7522

Table 4.1: PeMS loops inventory by Caltrans Districts.

detector stations in the PeMS database, where each station can have loops in as many as 10 lanes. More detector stations are added as Caltrans upgrades its infrastructure. Loop data from each Caltrans district are stored in a separate database table. This is because each district has its own TMC and collect data separately. In Los Angeles, for example, this table grows by one GB per day. The 30-second table contains the columns shown in Table 4.2.

ID	Timestamp	Lane	Occupancy	Flow
780062	02-aug-2002 15:37:20	2	0.0887	17
...				

Table 4.2: Raw table and example entries.

PeMS contains a configuration table with the definitions of the detectors, so data records can be referenced by detector ID. This table includes location information in freeway ID and postmile, detector type, number of lanes, etc. We call an individual loop detector in one lane a "loop," and a group of loops of the same type a Vehicle Detector Station (VDS). Figure 4.6 shows a typical configuration. The two loops marked "VDS_ID=3" in Figure 4.6 are in lanes 1 and 2 of the main line at the same location; the loop in the HOV lane at the same location belongs to a different VDS. The entries for VDS's $1, 2, 3, 4$ would appear in the configuration table as shown in Table 4.3.

Currently, loops provide all the traffic data in PeMS. However, PeMS is capable of collecting data from any source. Indeed, it is envisioned to collect and process all forms of transportation-related information. PeMS also collects incident data from the California Highway Patrol (CHP) website. The CHP posts incident information on a Web page. A

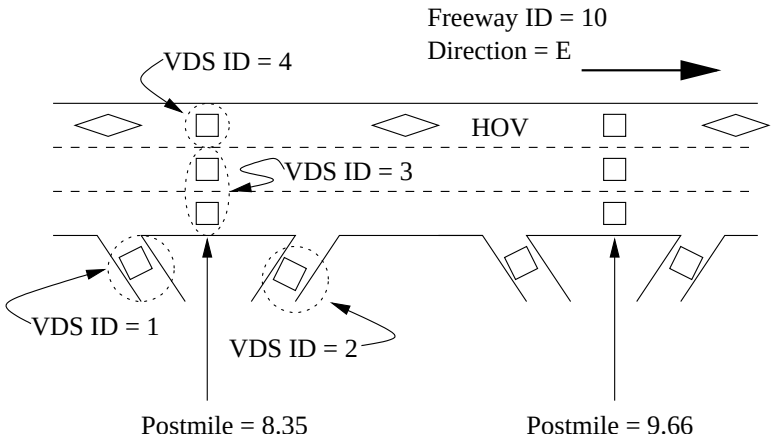


Figure 4.6: Loop locations and types

VDS ID	Freeway ID	Direction	Postmile	Type	No. of lanes
1	10	E	8.35	Main line	2
2	10	E	8.35	HOV	1
3	10	E	8.35	On ramp	1
4	10	E	8.35	Off ramp	1

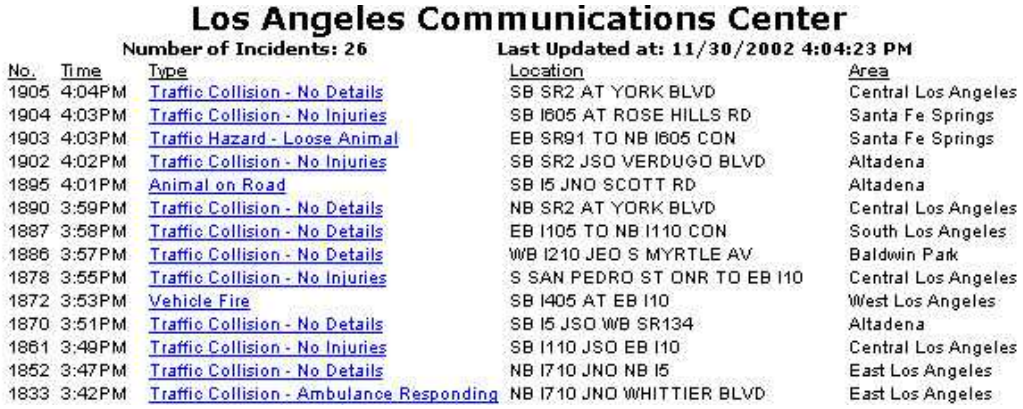
Table 4.3: Examples of loop configuration.

software agent accesses this page every 5 minutes, parses the HTML code, and stores the relevant information. Figure 4.7 shows a screen capture of the CHP website. This page is parsed to populate the database table shown in Table 4.4.

ID	From	To	Location	Type
1024921	05-aug-2002 08:30:00	05-aug-2002 09:42:00	SB US-101 JNO Willow	Vehicle on fire
...				

Table 4.4: Incident database table and sample entry.

Using the location information in Table 4.4 and Table 4.2, we can match the incidents to the affected detector loops for analysis. However, because they are in different



Los Angeles Communications Center

Number of Incidents: 26 **Last Updated at: 11/30/2002 4:04:23 PM**

No.	Time	Type	Location	Area
1905	4:04PM	Traffic Collision - No Details	SB SR2 AT YORK BLVD	Central Los Angeles
1904	4:03PM	Traffic Collision - No Injuries	SB I605 AT ROSE HILLS RD	Santa Fe Springs
1903	4:03PM	Traffic Hazard - Loose Animal	EB SR91 TO NB I605 CON	Santa Fe Springs
1902	4:02PM	Traffic Collision - No Injuries	SB SR2 JSO VERDUGO BLVD	Altadena
1895	4:01PM	Animal on Road	SB I5 JNO SCOTT RD	Altadena
1890	3:59PM	Traffic Collision - No Details	NB SR2 AT YORK BLVD	Central Los Angeles
1887	3:58PM	Traffic Collision - No Details	EB I105 TO NB I110 CON	South Los Angeles
1886	3:57PM	Traffic Collision - No Details	WB I210 JEO S MYRTLE AV	Baldwin Park
1878	3:55PM	Traffic Collision - No Injuries	S SAN PEDRO ST ONR TO EB I10	Central Los Angeles
1872	3:53PM	Vehicle Fire	SB I405 AT EB I10	West Los Angeles
1870	3:51PM	Traffic Collision - No Details	SB I5 JSO WB SR134	Altadena
1861	3:49PM	Traffic Collision - No Injuries	SB I110 JSO EB I10	Central Los Angeles
1852	3:47PM	Traffic Collision - No Details	NB I710 JNO NB I5	East Los Angeles
1833	3:42PM	Traffic Collision - Ambulance Responding	NB I710 JNO WHITTIER BLVD	East Los Angeles

Figure 4.7: CHP web page

forms – loops are specified by (freeway,direction,postmile) while CHP incidents are located by a string specifying (freeway,direction,cross street), we need to perform some processing before these different data types can be merged and used together. This 'data fusion' will be discussed in the next section on data processing.

PeMS collects weather information from the National Weather Service (NWS) in order to study the effects of rain, fog, and visibility on traffic. Similar to incident data, weather data are collected and stored by a software agent periodically.

Figure 4.8 shows the current PeMS data sources and data collection methods. Data are collected from Caltrans TMCs, the CHP website, and the NWS website. Because different TMC's publish data differently, PeMS uses several methods to collect loop data. For some districts, the data are sent by a server program running at the TMC and received by a client program at PeMS; for other districts, the TMC runs a program that directly inserts the data into the PeMS database.

Many other forms of data will become available in the future. For example, the

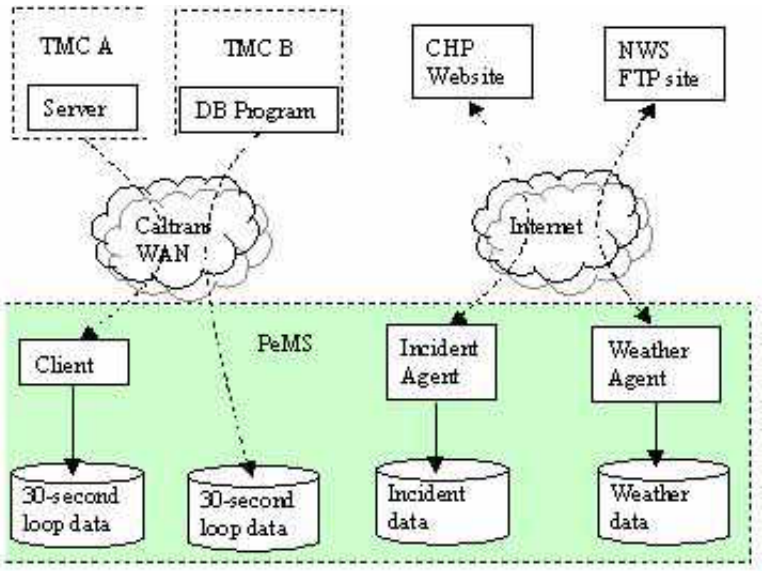


Figure 4.8: Data collection.

San Francisco Bay Area Metropolitan Transit Commission (MTC) is experimenting with a system to measure travel times from electronic toll tags. The direct measurement of travel times will improve the performance of PeMS's travel time calculations. The Berkeley Highway Laboratory has installed a video surveillance system that records and processes video images of traffic to generate individual vehicle trajectories. Data from non-freeway arterials and surface streets will augment those from existing freeway sensors. All these data are expected to be integrated into PeMS as they become available.

4.2 Data processing

The raw data need to be processed to become useful and informative. Suppose we want to find the delay yesterday on I-5N in Los Angeles. The raw data contain 30-second occupancy and volumes, but to find delay we need VMT and VHT. They, in turn, require

flow rate $Q(t, x)$ and speed $V(t, x)$ at times t at locations x along the freeway. Therefore, PeMS needs to estimate speed from occupancy and flow, computes VMT, VHT, delay, and travel time from speed and flow.

As this example shows, many interesting quantities are not directly measured and must be calculated from the raw data. Several important processing steps are listed below.

1. Compute derived values such as speed, VMT, VHT, delay, and travel time;
2. Diagnose data errors and impute missing values;
3. Aggregate data geographically and temporally;
4. Fuse data – relate data from different sources and in different formats.

We discuss each item in more detail below. The technical descriptions of the algorithms and their theoretical bases are given in Part II of this thesis.

4.2.1 Derived values

The 30 second data are converted into 5 minute average volume and occupancies, which are used to estimate 5-minute speed estimates using this formula:

$$\hat{V}(t) = \frac{Q(t)}{K(t)} L(t) \tag{4.1}$$

where $L(t)$ represents the average vehicle length of the time period t, at this location. Usually, $L(t)$ is taken to be a constant e.g. 20 feet. In reality, $L(t)$ can vary significantly among locations and times. $L(t)$ depends on the vehicle population at that time at that location, and also on the sensitivity of the detector. PeMS computes $L(t)$ using an algorithm described in Chapter 6.

4.2.2 Data quality

PeMS monitors the detector health. and replace bad and missing samples with imputed values using an algorithm. Data quality assurance is important in any sensor network. Samples are sometimes missing because of malfunctioning detectors or loss of communication. Some of the received samples are bad, meaning they are very unlikely to be the true values. Bad and missing data present problems when we want to compute traffic statistics. Bad values make the results of analysis suspect; missing samples complicate the analysis process because most algorithms operate on a regular set of data much more easily than those with potential missing samples. Therefore, detection of bad data and their imputation are important steps in the data flow. These algorithms are presented in Chapter 5.

4.2.3 Temporal and spatial aggregation

The large quantity of 30-second loop data need to be summarized to be interpretable. PeMS aggregates data at various levels in time and space. One may want the monthly congestion trend for the entire district, or summarize the performance of each freeway for comparison. It would be very inefficient to frequently compute these quantities from the raw loop data. Therefore, PeMS computes and stores aggregate measures to speed analysis. The quantities VMT, VHT, delay, and average speed are examples of aggregates. They are computed at the 5-min, 1-hour, and 1-day levels for each main line and HOV loop detector. Their computation is described in Chapter 2.

Travel time is another useful quantity that can be derived from the 5-minute

data. It is convenient to talk about travel time on segments, where a segment is a section of freeway between two nearest freeway crossings. In Los Angeles, there are 124 such segments. Segment travel times are computed for every 5-minute interval using the algorithm described in Section 3.2.2. They are used to quickly compute travel times on any route of multiple segments.

Figure 4.9 shows the different PeMS aggregates and their relevant time and space scales. For example, speed is meaningful for a 5-minute period in one location, but the average speed of all locations over a day is less meaningful; travel time for a certain starting time on a freeway or a segment is meaningful, but it's less meaningful to talk about it for a point location. PeMS stores these quantities at their natural time scales.

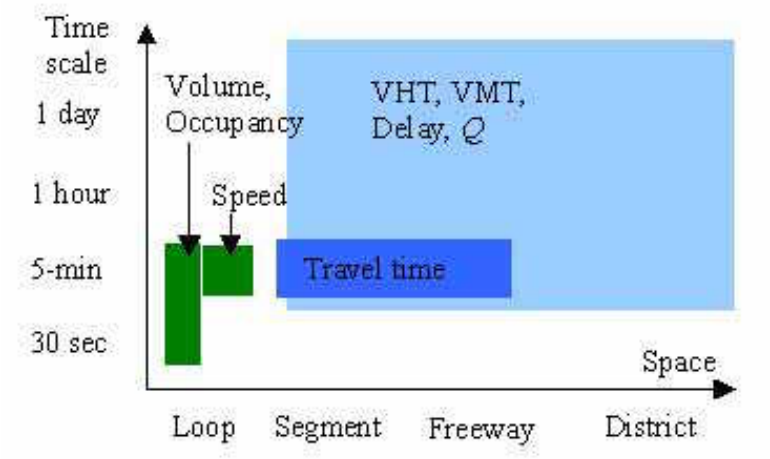


Figure 4.9: Quantities and their time and space.

The data processing procedures are invoked periodically at various time scales, from 30-sec to 5 minutes, to hourly, and daily. Figure 4.10 illustrates the data flow from raw data to aggregated. They are carried out by algorithms written in Perl and are fully

automated. The resulting database tables make it easy for higher level applications to display and analyze data.

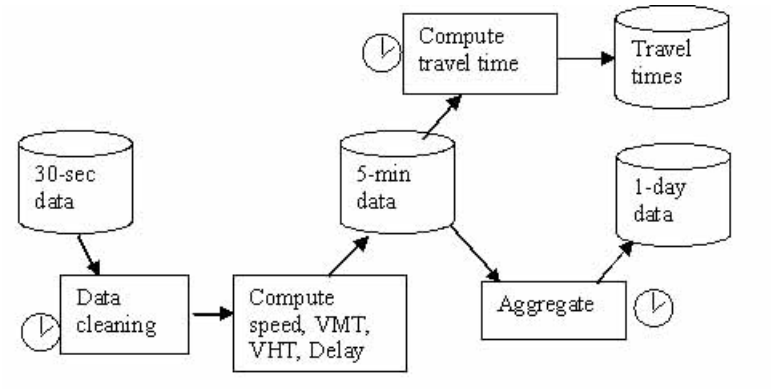


Figure 4.10: Data flow.

4.2.4 Data fusion

When we studied the effect of incidents on traffic in Sections 3.2 and 3.3, we needed to study loop data and incident data together. These data come from different sources and in different formats, and combining them is an example of data fusion.

From loop detectors, we obtain measurements of flow and occupancy, and we calculate speed, for sample times t_i at detector locations x_i. The location is a scalar, representing the postmile on the freeway. On the same freeway and in the same direction, we also obtain incident data, in the form of Table 4.4. In this form, the incidents cannot be directly related to loop data. We first have to convert the location from a text string representing the freeway and cross street to a scalar representing the postmile. This is done using a Geographical Information System (GIS) software, ArcView. The incidents can then be represented in a way similar to loop data. Let $I(t, x)$ be the incident type at (t, x),

such that $I(t, x) \in 0, 1, 2, \ldots, n$. In this formulation, $I = 0$ if there is no incident at (t, x); if there is an incident, its type is enumerated as an integer between 1 and n for n possible types. In Section 3.3, for example, I takes on values 0, acc, and non, which represent no incident, accident, and non-accident incident.

As we obtain other forms of data, we will need to fuse more types of data together. For example, we map stations to detector locations in order to assign weather conditions to freeway locations. The value of PeMS data grows with each additional source and the fusion of different data types.

4.3 Web and other access methods

The primary way to access PeMS is via the World Wide Web. Web browsers are ubiquitous, and most people are familiar with web pages and their control and navigation. There are many tools for developing web pages and user interfaces. Using a web interface allows us to leverage the developments in this popular technology.

PeMS uses the Apache web server. Most of the applications described in Chapter 2 are written in PHP. Software and applications were developed by Karl Petty and his company, Berkeley Transportation Systems (BTS), a subcontractor to PeMS. PHP, for Personal Home Page, is a scripting language designed for the web. Applications written in PHP are invoked by the user and executed by the web server. They contain a small amount of logic to serve user requests, query the database, and generate formatted outputs such as tables and graphs.

The applications query the processed database tables rather than the raw tables

for speed and simplicity. Data processing in PeMS are handled by special algorithms that run in the background and populate the processed tables. Therefore, the web applications don't need to contain a lot of logic. They simply query the quantities that have already been computed. An example of an application is the one that plots delay for a district over a year, shown in Figure 2.3. Delay is calculated from flow and speed of each detector for every 5 minutes. But when this plot is requested, the delays for the district have already been calculated and stored by the periodic data processing steps. The web application needs only to query the processed table that contains delay for the entire district on each day. This is a much smaller table than the raw data table, so the application can run very quickly. Our plots are generated by GNU Plot[1], which is invoked by the PHP applications, as indicated in Figure 4.11.

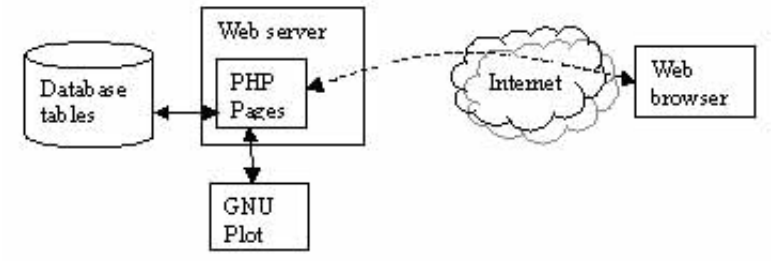


Figure 4.11: Web access architecture.

Sometimes, the web applications do not satisfy the needs of very specific analyses, such as the ones presented in Chapter 3. For these studies, we wrote custom applications to directly access the database. These applications can be written in any language, and we mostly use Perl, Matlab, Java, or SQL. Accessing the database directly requires knowledge of the tables and the data.

[1]GNU Plot is a free program that plots data from a file. The output is sent to an image file, like GIF.

There are many other ways to access PeMS. By storing all of our data on-line, we hope to promote their use by traffic managers and researchers. We also designed the system to be open and easily maintained and upgraded. This encourages the further development of PeMS and additions to its capabilities.

Part II

Theory And Algorithms

Chapter 5

Data Quality

We found that while the enormous amount of loop and incident data contain a wealth of information, extracting it from the raw data requires sophisticated algorithms based on traffic theory and statistics. In the second part of this thesis, we present several important algorithms in detail.

Detection and imputation of data errors are crucial first steps in the data flow. Bad and missing samples present problems for any algorithm that use the data for analysis. PeMS's detection and imputation algorithms produce a complete grid of values from the raw data and simplifies the design and operation of higher level algorithms.

We need to detect bad data from the measurements themselves, because they are the only quantities we have. There has been some work in this area by the FHWA, Washington DOT, and others. Existing algorithms usually operate on the raw 20-second or 30-second level, and produce a diagnosis for each sample. But it's very hard to tell if a single 20-second sample is invalid. Fortunately, loop errors are not completely random.

Some detectors report good data most of the time, while others seem to produce errors most of the time. We find that it's possible to identify most errors from time series plots of measurements.

PeMS detects malfunctioning detectors based on the previous day's data. This algorithm reliably finds misbehaving loops. There is a small fraction (about 2%) that produce errors intermittently. These data are detected by PeMS's real time detection algorithm that diagnoses each 5-minute sample. It uses information from the previous day and neighbor measurements.

Erroneous and missing samples leave holes in the data that must be filled with imputed values. Imputation using time series analysis has been suggested in previous works, but these methods are only effective for short periods of missing data; linear interpolation and neighborhood averages are natural imputation methods, but they don't use all the data available. We developed an imputation algorithm that uses data from neighbor detectors. This algorithm models a linearly relationship between each pair of neighbors, and fits its parameters on historical data. It is robust, and is shown to perform better than other methods.

5.1 Notation

In this and the following chapters, we use capital letters Q and K to denote flow rate and occupancy.

5.2 Existing detection methods

The poor quality of loop data has plagued their effective use for a long time. In 1976, Payne [33] identified five types of detector errors and presented several ways to detect them from 20-second and 5-minute volume and occupancy measurements. These methods place thresholds on minimum and maximum flow, density, and speed, and declare a sample to be invalid if it fails any of the tests. Following this work, Jacobsen and Nihan at the University of Washington defined an acceptable region in the K-Q plane, and declared samples to be good only if they fell inside [34]. We call this the Washington Algorithm. The boundaries of the acceptable region are defined by a set of parameters which either need to be calibrated on historical data, or derived from traffic theory.

Existing detection algorithms [34] [33] [35] [36] try to detect the errors described in [33]. For example, chattering and pulse break up cause Q to be high, so a threshold on Q can catch these errors. But some errors cannot be caught this way, such as a detector stuck in the "off" or (Q=0,K=0) position. Payne's algorithm identifies this as a bad point, but good detectors also report (0,0) when there are no vehicles in the detection period as can happen during periods of very light traffic. Therefore, eliminating all (0,0) points introduces a positive bias in the data. On the other hand, the Washington Algorithm accepts this point, but doing so makes it unable to detect the "stuck" type of error. Similarly, a threshold on occupancy is also hard to set. An occupancy of 0.5 for one 30-second period does not necessarily indicate an error, but a large number of 30-second samples with occupancies of 0.5, especially during non-rush hours, points to a malfunction.

We implemented the Washington Algorithm using default parameters and applied

it to one day of 30-second data from 2 loops in Los Angeles. The region of acceptability was taken from [34]. The data and diagnoses are shown in Figure 5.1. The first row shows

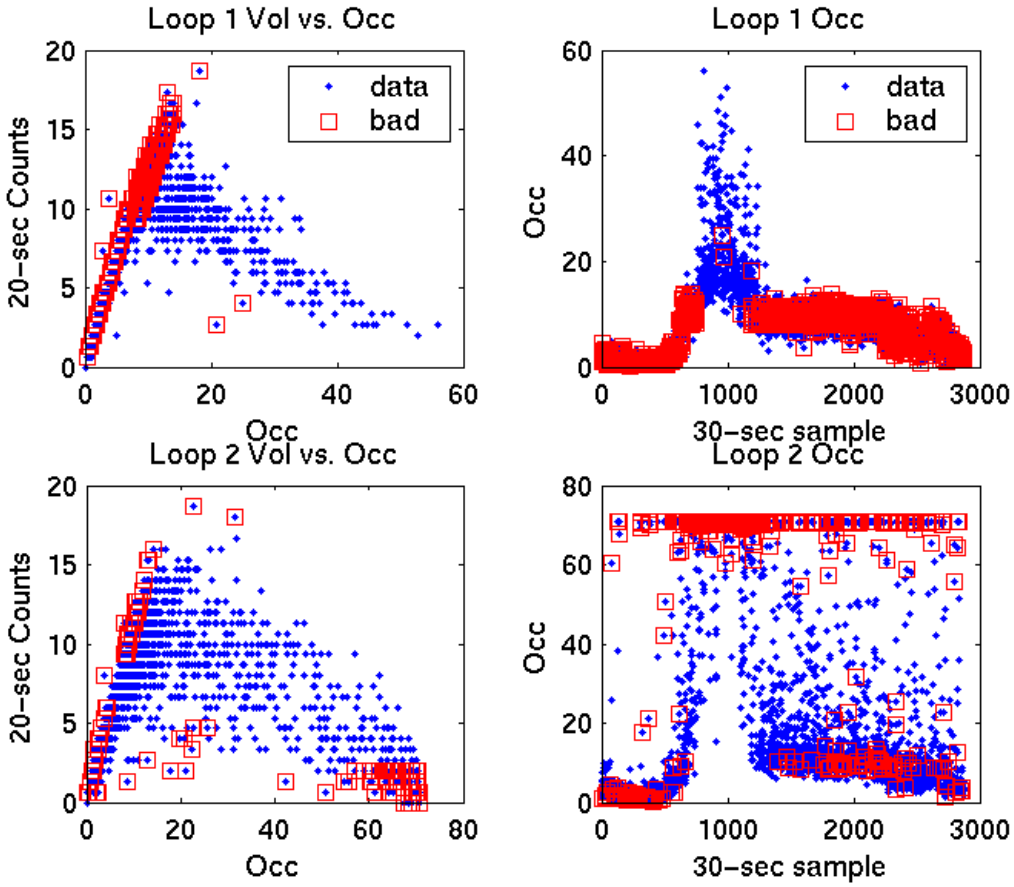


Figure 5.1: Washington Algorithm applied to Los Angeles Data.

loop 1. On the left, sample flows are plotted against occupancy. The samples declared as bad are marked by squares. The Washington Algorithm flagged many samples on the left (with high ratio of Q/K) while they appear to be good. This suggests that the parameters of the algorithm need to be set for data from our location. The second row shows loop 2, which appears to be bad because many occupancy samples are very high, even during

periods when traffic light in the next lane. The Washington Algorithm failed to capture many of the bad samples. This is because the region of acceptability has to be very large to not rule out possibly valid observations. However, because we easily recognize loop 1 to be good and loop 2 to be bad by looking at the time series plots of their occupancies, a detection algorithm should consider more than one sample at a time. This is the key insight that led to the PeMS loop error detection algorithm.

5.3 PeMS daily detection algorithm

5.3.1 Design

We present an algorithm for loop error detection that uses the time series of flow and occupancy measurements, instead of making an independent detection for each individual sample. It is based on the observation that good and bad detectors behave very differently over time. For example, at any given instant, the flow and occupancy at a detector location can have a wide range of values, and one cannot rule most of them out; but over a day, most detectors show a similar pattern - flow and occupancy are high in the rush hours and low late at night.

Figure 5.2 shows typical 30-second flow and occupancy measurements. Most loops report samples that look like this, but some loops behave very differently. Figure 5.3 shows an example of a bad loop. This loop has zero flow and occupancy of 0.7 for several hours during the evening rush hour - clearly, this cannot be the true traffic behavior. We found four types of abnormal time series behavior, and list them in Table 5.1. Types 1 and 4 are self-explanatory; types 2 and 3 are illustrated in Figures 5.1 and 5.3. The errors in Table 5.1

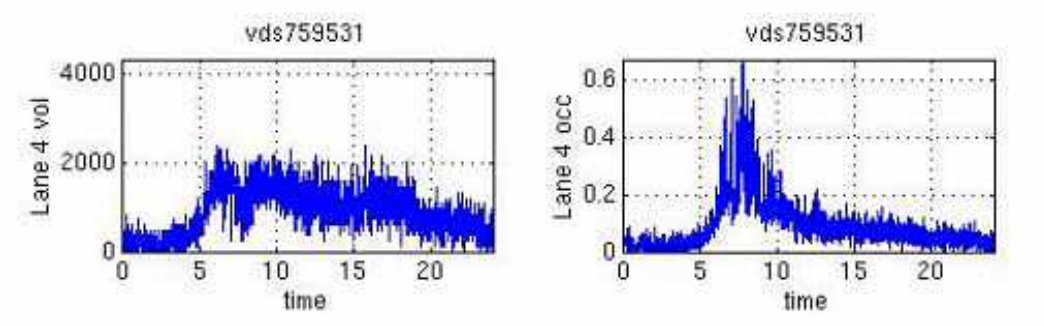


Figure 5.2: Example of a good loop. Data from I-5N in Los Angeles at mainline postmile 8.27 on 8/7/2001.

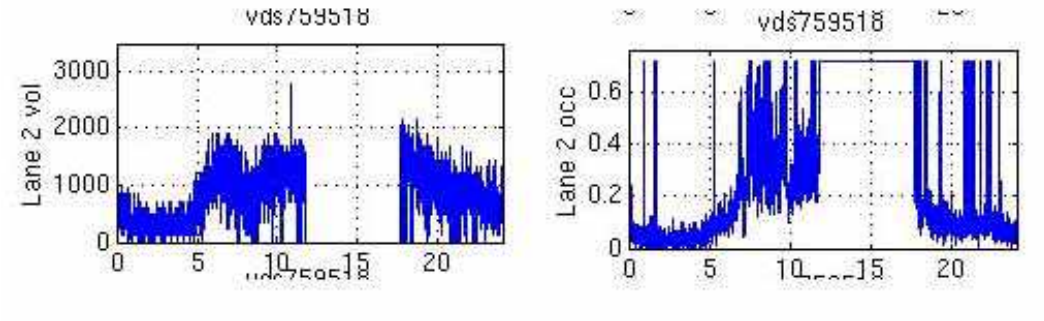


Figure 5.3: Example of a bad loop. Data from I-5N at mainline postmile 4.58 on 8/7/2002.

are not mutually exclusive. For example, if a loop has all zero occupancies, then it belongs to both type 1 and type 4. A loop is declared bad if it is in any of these categories.

Error Type	Description	Likely Cause	Percent
1	Occupancy and flow are zero	Stuck off	5.6%
2	Non-zero occupancy and zero flow, example in Figure 5.3	Hanging-on	5.5%
3	Very high occupancy, example in Figure 5.1	Hanging on	9.6%
4	Constant occupancy and flow	Stuck on or off	11.2%
All Errors			16%

Table 5.1: Error Types

We did not find a significant number of loops that have chatter or pulse break up,

which would produce abnormally high volumes, therefore we don't check for this condition in the current version of the detection algorithm. However, a fifth error type and error check can easily be added to this scheme to flag loops with consistently high counts.

Let

$$\Delta_i(d) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if the } i\text{th loop is bad on the } d\text{th day,} \\ 0 & \text{otherwise.} \end{cases}$$

This is an unknown condition, which we will try to estimate using data from day d. We developed the Daily Statistics Algorithm (DSA) to recognize error types 1-4 above. The input to the algorithm is the time series of 30-second measurements $Q(d,t)$ and $K(d,t)$, where d is the index of the day, and $t = 0, 1, 2, \ldots, 2879$ is the 30-second sample number; the output is the estimate of diagnosis $\hat{\Delta}_i(d)$. In contrast to existing algorithms which operate on each sample, the DSA produces one diagnosis for each loop on each day. We base the diagnosis on only samples between 5am and 10pm, because outside of this period, the traffic is light and it's more difficult to tell the difference between good and bad loops. There are 2041 30-second samples in this period, therefore the algorithm is a function of 2041x2=4082 variables,

$$\hat{\Delta}_i(d) = f(\{Q(d,t), K(d,t) : \; t_0 \leq t \leq t_1\})$$

where $t_0 = 5 \times 120$ and $t_1 = 22 \times 120$ are the limits of the diagnostic period. These large number of observations are summarized by four statistics $S_, \ldots, S_4$ in Table 5.2, where $S_j(i,d)$ is the jth statistic computed for the ith loop on the dth day. The decision $\hat{\Delta}$ becomes a function of these four variables. The diagnosis is straightforward: a loop is declared bad if any of the variables exceeds their thresholds $s_1^*, \ldots, s_4^*$.

Name	Definition	Description
$S_1(i,d)$	$\sum_{a\le t\le b} 1(K_i(d,t)=0)$	Number of samples with zero occupancy
$S_2(i,d)$	$\sum_{a\le t\le b} 1(K_i(d,t)>0)1(q_i(d,t)=0)$	Number of samples with zero flow and positive occupancy
$S_3(i,d)$	$\sum_{a\le t\le b} 1(K_i(d,t)>k^*),\ k^*=0.35$	Number of samples with occupancy greater than k^*
$S_4(i,d)$	$(-1)\sum_{x:\hat{p}(x)>0}\hat{p}(x)\log(\hat{p}(x))$ $\hat{p}(x)=\frac{\sum_{a\le t\le b}1(K_i(d,t)=x)}{(b-a+1)}$	Entropy of occupancy samples

Table 5.2: Statistics computed from measurements.

$$\hat{\Delta}_i(d) = 1 - 1(S_1(i,d) > s_1^*)1(S_2(i,d) > s_2^*)1(S_3(i,d) > s_3^*)1(S_4(i,d) > s_4^*) \tag{5.1}$$

These four statistics were chosen because they are good indicators of the four types of loop failure listed in Table 5.1. To illustrate, we show the histogram of each of $S_1, \ldots, S_4$ in figures 5.4 - 5.7. The data are collected from 5377 loops in Los Angeles on 4/24/2002.

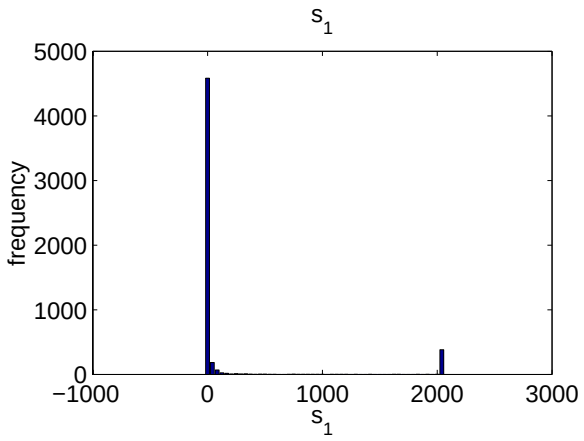


Figure 5.4: Distribution of S_1.

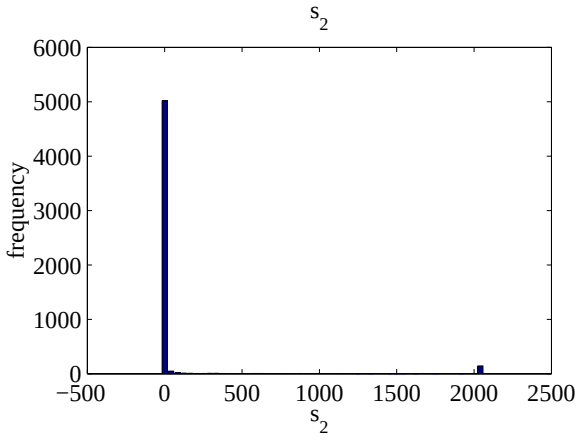


Figure 5.5: Distribution of S_2.

The distribution of each of the statistics shows two distinct populations. In S_1, for example,

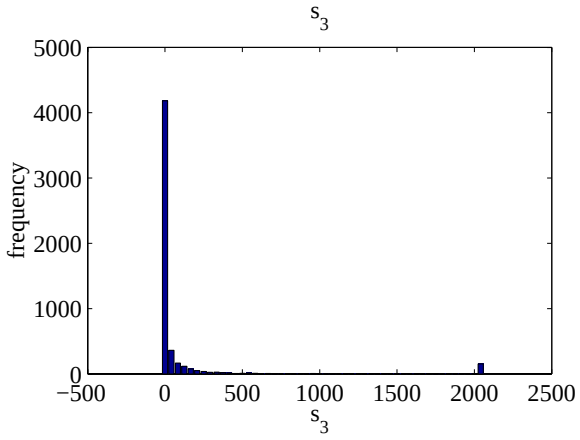


Figure 5.6: Distribution of S_3.

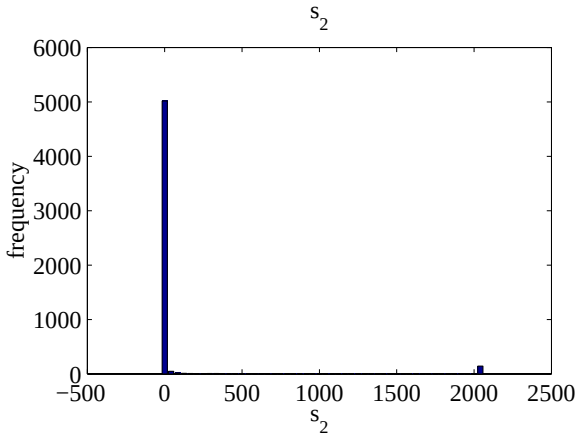


Figure 5.7: Distribution of S_4.

there are two peaks at $S_1 = 0$ and $S_1 = 2041$. This shows that there are two groups of loops, one group of about 4700 loops has very few samples that show zero occupancy, while another group of about 300 has almost all zeros. The second group is bad, because they have type 1 error. Since all the distributions are strongly bimodal, (5.1) is not very sensitive to s_j^*. We only need to be able to separate the two peaks in each of the four histograms in figures 5.4 - 5.7. The default thresholds are given in Table 5.3. The only other parameters of this model are the time ranges, and the definition of S_3, where an occupancy threshold of 0.35 is specified. The DSA uses a total of seven parameters; they work well in all six Caltrans districts.

5.3.2 Performance

The daily detection algorithm was implemented and tested on PeMS data. The last column in Table 5.1 shows the distribution of the four types of errors for 3039 loops in District 12 (Orange County) for 31 days in October, 2001. Because we don't have the

Parameter	Value
k^*	0.35
s_1^*	1200
s_2^*	50
s_3^*	200
s_4^*	4
a	5am
b	10pm

Table 5.3: Default parameters

ground truth of actual Δ_i's, , we have to verify the performance of this algorithm visually. Fortunately, this is easy in most cases, because the time series show distinctly different patterns for good and bad detectors.

We test the algorithm Orange County on 10/1/2001. Figure 5.8 shows 10 randomly selected loops from those that are declared bad. Plot 1 appears to show good data. It was diagnosed as bad because its entropy was slightly below the threshold. The algorithm make correct diagnoses on the other nine loops. Plots 4 through 8 have all zero occupancies; plot 2 has mostly zeros; plots 3 and 9 have many samples with very high occupancies at all times of the day; plot 10 looks mostly normal, but has some values of high occupancies where there was zero flow (not shown).

Figure 5.9 shows 10 loops randomly selected from the those declared to be good. They all appear to be good.

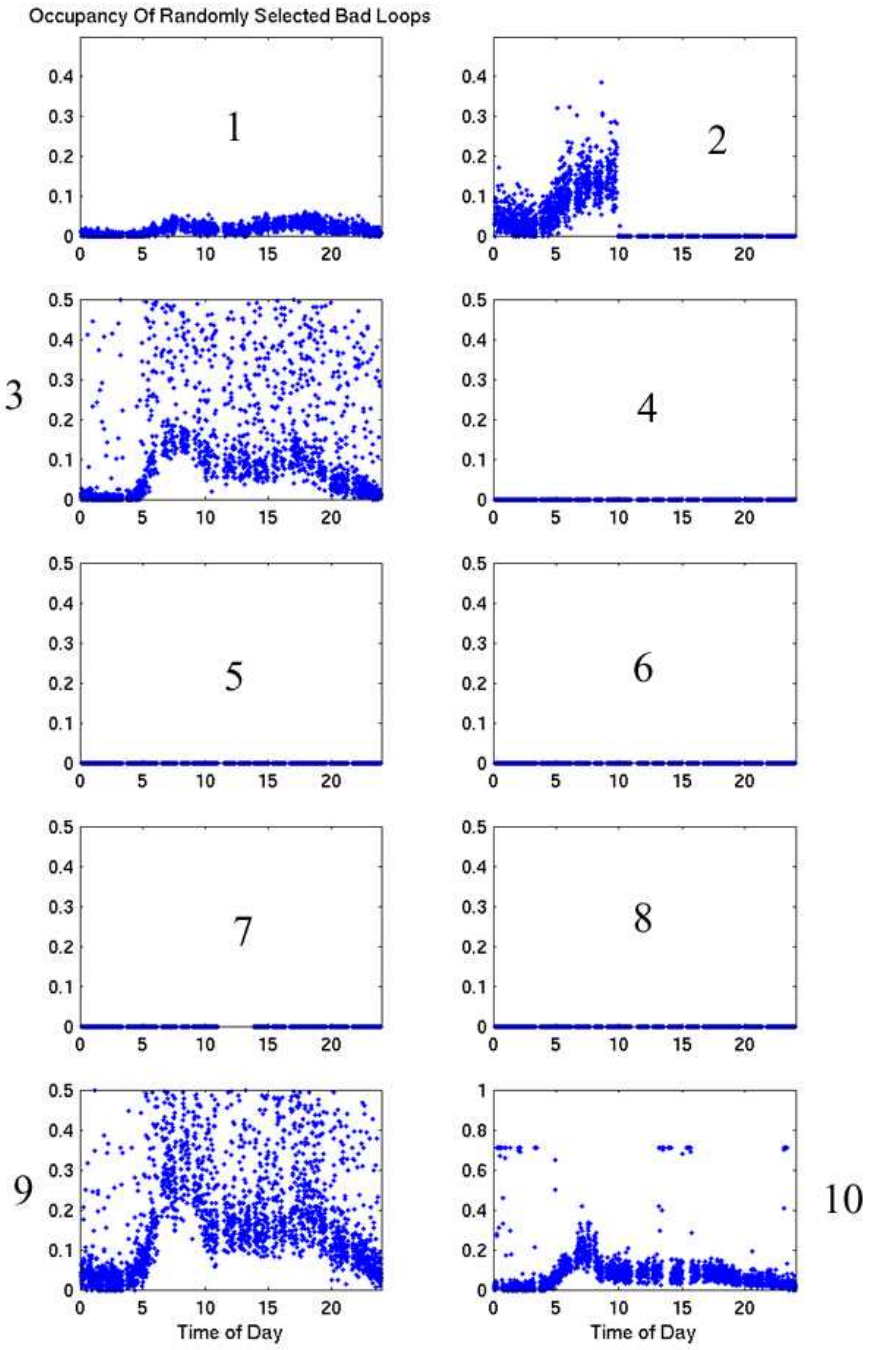


Figure 5.8: Loops declared as bad.

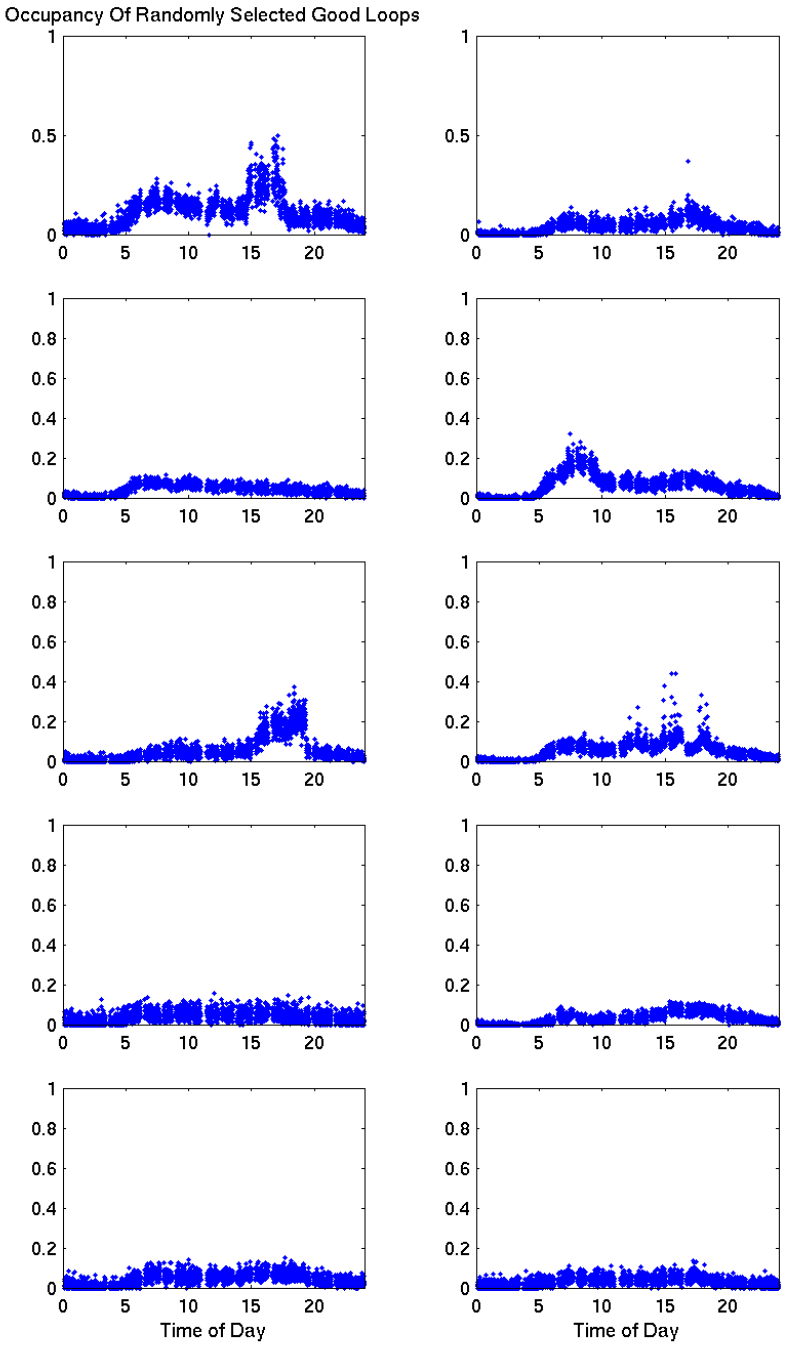


Figure 5.9: Loops declared as good.

A larger visual test was performed on loops in Los Angeles, on data from 8/7/2001. There are 662 loops on Interstate 5 and Interstate 210, out of which 142 (21%) were declared to be bad by the algorithm. We then manually checked the plots of occupancy to verify these results. We found 14 loops that were declared good to be inconclusive from a visual inspection, suggesting a possible underdetection of 14/(662-142) = 2.7%. There were no false positives. These results show that the algorithm performs reliably.

$\hat{\Delta}_i(d)$	Looks good	Looks bad	Total
1	0	142	142
0	496	14	510
Total	496	156	662

Table 5.4: Summary of visual test.

5.3.3 Real-time approximation

The detection algorithm described above identifies most bad loops because they tend to remain bad, or remain good, for a long time. Although the diagnostics has a lag of one day, the previous day's diagnosis accurately predicts the current day's results. Table 5.5 shows the conditional probabilities of $\hat{Delta}_i(d+1)$ given $\hat{Delta}_i(d)$ based on data from Orange County. Overall, the probability that a loop's diagnosis changes from good to bad, $P(\Delta_i(d-1) \neq \Delta_i(d))$, is 2%. For the small fraction of loops that produce intermittent

δ	$P(\hat{\Delta}_i(d) = \delta)$	$P(\hat{\Delta}_i(d) = \delta \mid \hat{\Delta}_i(d-1) = \delta)$
0	84%	99%
1	16%	94%

Table 5.5: Accuracy of next-day prediction

errors, we developed a true real time detection scheme that acts on each sample. This algorithm is discussed in Section 5.5.

5.4 Imputation

5.4.1 The need for imputation

We model the measurement of each detector as either the actual value or an error value, depending on the state of the detector Δ_i. Let $K_i(d,t), Q_i(d,t)$ be the true occupancy and flow, and let $\tilde{K}_i(d,t), \tilde{Q}_i(d,t)$ be the measured values. The measurement represents the true value only if the detector is good:

$$\Delta_i(d) = 0 \Rightarrow \begin{cases} \tilde{K}_i(d,t) = K_i(d,t), \\ \tilde{Q}_i(d,t) = Q_i(d,t). \end{cases} \tag{5.2}$$

When $\Delta_i(d) = 1$, $\tilde{K}$ and $\tilde{Q}$ are independent of the true occupancy and volume. That is, the measurements tell us nothing about the true values if the detector is bad.

The model above implies that we should discard the measurements if $\hat{\Delta}_i(d) = 1$. Doing so creates holes in additional to originally missing samples, and the presence of holes complicates the design of data analysis algorithms. The holes must be filled with estimates. Missing data can be filled using time series analysis. Nihan modeled occupancy and flow time series as ARMA processes and predicted values in the near future [37]; Dailey presented a method of prediction from neighbor loops using a Kalman filter [38]. In PeMS, loops errors do not occur randomly, but persist for many hours and days. Time series predictions like those of [37], [38] become invalid very quickly and cannot be used in these situations.

We developed an imputation scheme that uses information from good neighbor

loops at the current sample time. This is a natural way of dealing with missing data. Traditional techniques for filling in missing data sometimes implicitly use neighbor values. For example, at a location that has four lanes but only three of them have measurement, the total flow rate can be estimated as the average flow rate of the three lanes multiplied by four. This method implicitly imputes the missing value as the average of its neighbors. Linear interpolation is an example of explicit imputation. While these traditional imputation methods are intuitive, they make naive assumptions about the data. The PeMS imputation algorithm, on the other hand, models the behavior of neighbor loops using their historically observed relationship.

5.4.2 Linear model of neighbor detectors

We find that occupancies and volumes of detectors in nearby locations are highly correlated. Therefore, measurements from one location can be used to estimate quantities on the others, and a more accurate estimate can be formed if all the neighboring loops are used in the estimation. We define two loops to be neighbors if they are in the same location in different lanes, or if they are in adjacent locations. Figure 5.10 shows a typical neighborhood. Examples of linear dependencies of neighbors are shown in Figures 5.11 and 5.12. Figure 5.13 plots the distribution of the correlation coefficients between all neighbors in Los Angeles. It shows that most neighbor pairs have high correlations in both flow and occupancy.

The linear relationship between neighbor loop measurements allows the use of linear regression to estimate values at a loop from the measurements of its neighbors. We

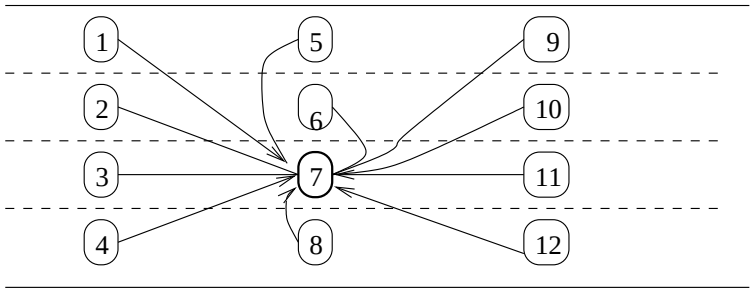


Figure 5.10: Loops and their neighbors.

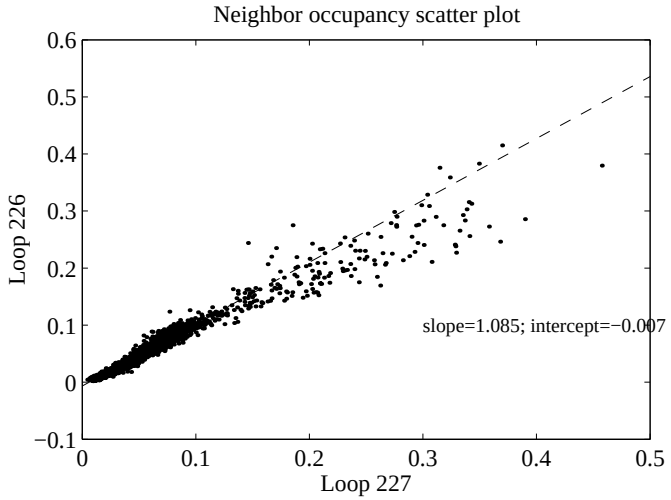


Figure 5.11: Linear relationships of occupancies of two neighbors.

use the following linear model to relate the measurements from neighbor loops i and j:

$$Q_i(d,t) = \alpha_0(i,j) + \alpha_1(i,j)Q_j(d,t) + \epsilon_{i,j}(d,t), \tag{5.3}$$

$$K_i(d,t) = \beta_0(i,j) + \beta_1(i,j)K_j(d,t) + \xi_{i,j}(d,t), \tag{5.4}$$

where ϵ and ξ are Gaussian noise. While the intercepts $\alpha_0(i,j)$ and $\beta_0(i,j)$ are allowed to be negative, the quantities Q and K themselves are always non-negative. When using this model to estimate neighbor quantities, we force the estimates to be also non-negative.

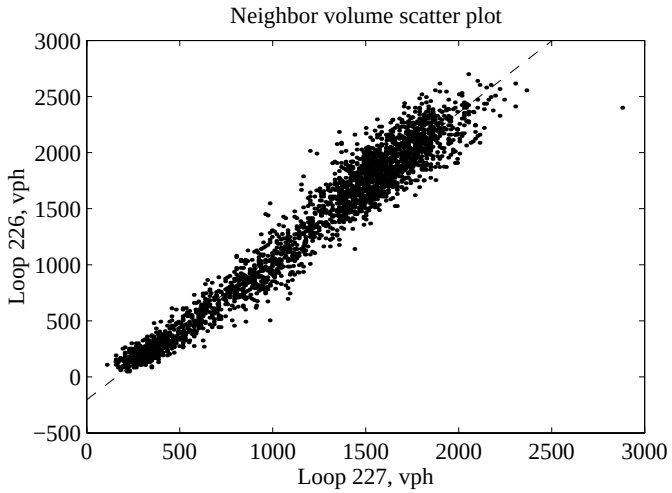


Figure 5.12: Linear relationships of volumes of two neighbors.

For each pair of neighbors (i, j), the parameters $\alpha_0(i, j)$, $\alpha_1(i, j)$, $\beta_0(i, j)$, $\beta_1(i, j)$ are estimated using several days of historical data. We use only those loops that were diagnosed as good on those days.

$$\begin{bmatrix} \hat{\alpha_0} \\ \hat{\alpha_1} \end{bmatrix} (i, j) = \underset{(\alpha_0, \alpha_1)^T}{\arg\min} \left(\frac{1}{n_t n_d} \sum_{d:\, \hat{\Delta}_i(d) = \hat{\Delta}_j(d) = 0} \sum_{t} \left[\tilde{Q}_i(d, t) - \alpha_0 - \alpha_1 \tilde{Q}_j(d, t) \right]^2 \right). \quad (5.5)$$

In the above, n_t is the number of samples of each day, n_d is number of days of training data. The parameters (β_0, β_1) for K are fitted the same way. In Los Angeles, there are 60,700 neighbor pairs. Using data from 10 days between 4/11/2002 and 4/20/2002, we found parameters for neighbor pairs when both loops have data. There were 34,684 such pairs. Figure 5.14 shows the distributions of slopes $\alpha_1(i, j)$ and $\beta_1(i, j)$. They are distributed near 1. The median $\alpha_1(i, j)$ is 0.94 and the median $\beta_1(i, j)$ is 0.88.

Using (5.5), we can find parameters for neighbors (i, j) as long as there are historical data from those detectors. But some detectors have no historical data. This can

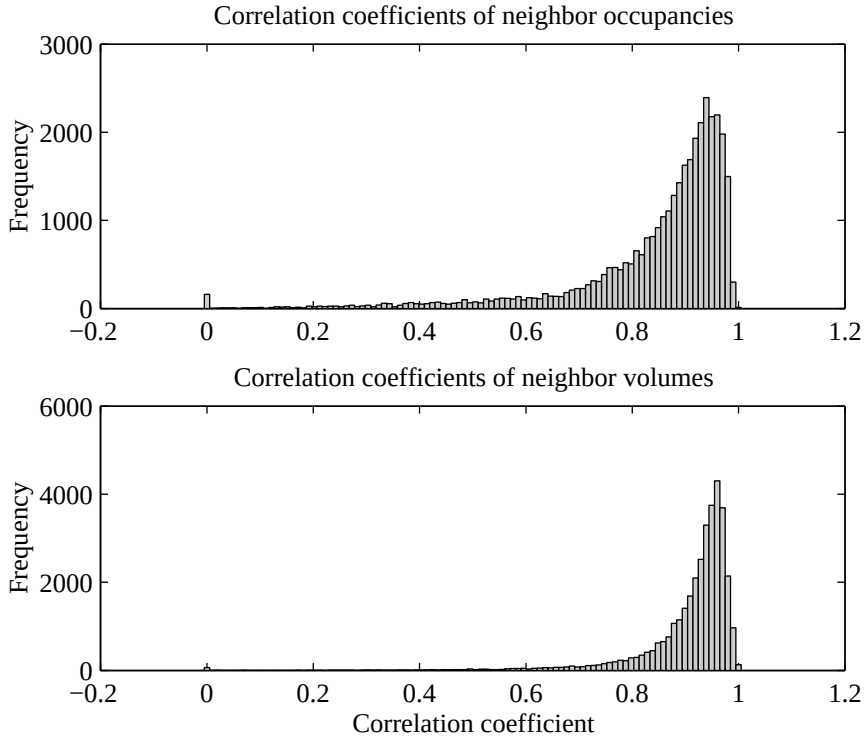


Figure 5.13: Distribution of correlation coefficients.

happen when a location has four lanes, where only lanes 2, 3, and 4 have historical data but not lane 1. We cannot find imputation parameters for any neighbor pair containing lane 1. In these cases, we estimate regression parameters based on the behavior of similar locations where there are historical data. We call parameters estimated this way the global parameters.

We expect similarly related pairs of loops to behave similarly. For example, flow rate is generally highest in lane 1, then lane 2, and so on. We use three configuration parameters to specify the type of neighbor pair, δ_{ij}, l_i, l_j, where

$$\delta_{ij} = \begin{cases} 0 & \text{if } i,j \text{ are in the same location} \\ 1 & \text{otherwise} \end{cases}$$

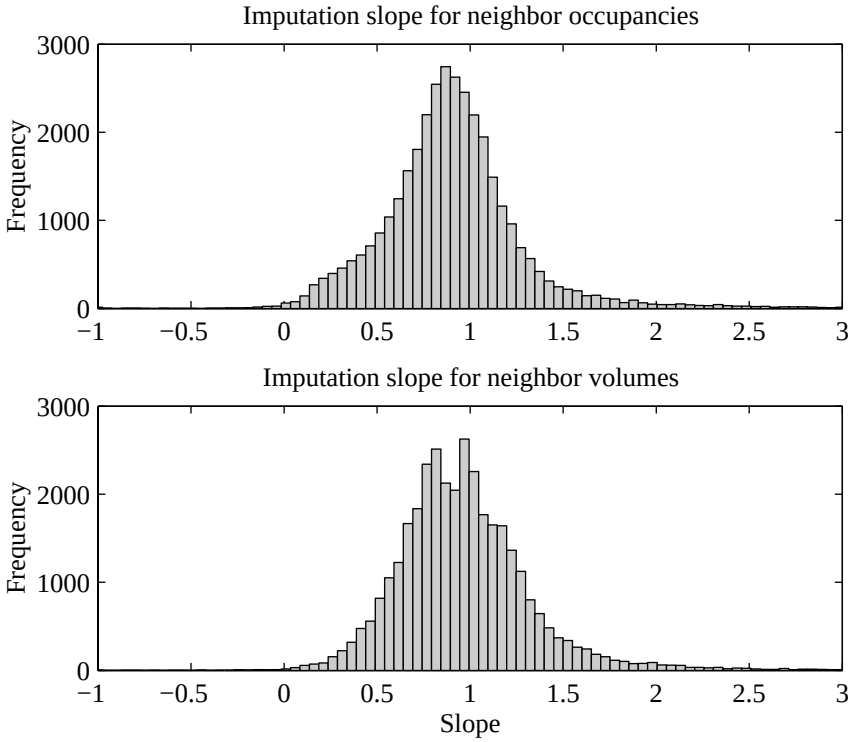


Figure 5.14: Distribution of imputation slope for neighboring occupancy and volume in Los Angeles.

$$l_i, l_j \quad = \quad 1, 2, \ldots = \text{lane numbers of loops } i \text{ and } j$$

For each configuration (δ, l, l'), we obtain linear coefficients $\hat{\alpha}_0(\delta, l, l')$, $\hat{\alpha}_1(\delta, l, l')$. using

$$\begin{bmatrix} \hat{\alpha_0} \\ \hat{\alpha_1} \end{bmatrix} (\delta, l, l') = \underset{(\alpha_0, \alpha_1)^T}{\arg\min} \left(\frac{1}{n_t n_d} \sum_{\substack{\{i,j:\, \delta_{ij}=\delta, l_i=l, l_j=l'\}, \\ \{d:\, \hat{\Delta}_i(d)=\hat{\Delta}_j(d)=0\}, \\ \{t\}}} \left[\tilde{Q}_i(d,t) - \alpha_0 - \alpha_1 \tilde{Q}_j(d,t) \right]^2 \right) \tag{5.6}$$

The global parameters for K are similarly computed.

Each neighbor j in the neighborhood of i contributes an estimate for the value of i, and the final imputation result is the median of the pairwise estimates,

$$\hat{K}_{ij}(d,t) \quad = \quad \hat{\beta}_0(i,j) + \hat{\beta}_1(i,j) \tilde{K}_j(d,t), \tag{5.7}$$

$$\hat{K}_i(d,t) \quad = \quad \underset{\left\{j \in A_i,\, \hat{\Delta}_j(d)=0\right\}}{\text{median}} \hat{K}_{ij}(d,t), \tag{5.8}$$

where A_i is the neighborhood of i. Both volume and occupancy are imputed the same way, and only estimates from good neighbors are used in the imputation. Equation (5.8) robustly combines information from multiple neighbors. It was first suggested by Professor John Rice of the Statistics Department at UC Berkeley. An alternate estimation method is multiple regression, in which the estimate is a linear combination of all the neighbors, rather than the two step process of (5.7) and (5.8). While multiple regression may perform better when all neighbors are available, the pairwise method is more robust because it produces a valid estimate for any combination of good/bad neighbors as long as there is at least one good neighbor. We require robustness because of the frequency of errors. Multiple regression requires a different set of coefficients for each combination of good/bad neighbors. Dailey also presented an imputation method based on all neighbors jointly using a Kalman filter [38]. This method is similar to multiple regression. Robustness is also the reason for choosing the median in (5.8) instead of the mean, which is affected by outliers and errors in $\hat{Delta}_j$.

The above imputation scheme produces an estimate for any loop that has at least one good neighbor, but loops that have no good neighbors are not imputed. However, further imputation can be performed by treating imputed loops as good and using their imputed values to estimate the values at their neighbors. With each iteration, the set of "good" loops grows until no more imputation can be performed, all the loops have been imputed, or a maximum number of iterations is reached. There needs to be a maximum allowed number of iterations because the accuracy of the estimates decreases with each

iteration, and at some point another imputation method may be better for loops that still don't have estimates. In our experience, most of the bad loops are filled after the first iteration. In District 7 on 4/24/2002, for example, the percentages of loops filled in the first four iterations are 90%, 5%, 1%, 1%; the entire grid is filled after eight iterations. After the maximum number of allowed iterations, the loops that are still not imputed are simply imputed with their historical averages at the same time of day.

5.4.3 Performance

We evaluated the performance of the imputation algorithm on data from 4/24/2002. To run this test, we found 189 loops that were themselves good and also had good neighbors. From each loop i, we collected the measured flow and occupancy $\tilde{Q}_i(t)$ and $\tilde{K}_i(t)$. We then ran the algorithm to compute their estimated values, and found the root mean squared error for each loop. See Table 5.6. This table shows that the estimates are unbiased as

Quantity	Mean	Standard deviation	Mean Absolute Error	Standard deviation of error	Mean Error
Occupancy (no unit)	0.085	0.061	0.013	0.021	0.0001
Volume(vph)	1220	527	132	201	6

Table 5.6: Imputation performance.

they should be. The standard deviation of imputation error is small compared to the mean and standard deviation of the measurements. Figure 5.15 shows the comparison between estimated and original values for one loop as an example. They show good agreement.

We also compared the performance of our algorithm against that of linear inter-

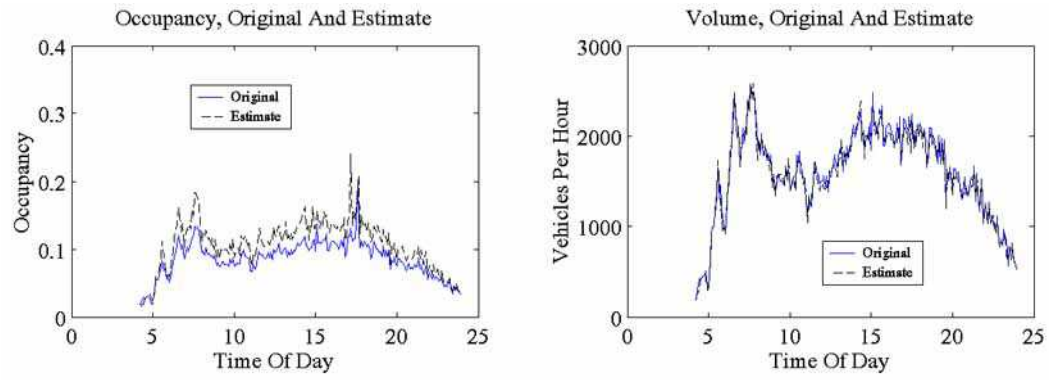


Figure 5.15: Comparison between imputed and actual values of occupancy and flow.

polation. Fifteen triplets of good loops were chosen for this test. Ten of the triplets are loops in the same lane at different locations, while five other triplets have their loops in the same location, across three lanes. In each triplet, we used two loops to predict the volume and occupancy of the third loop using linear interpolation. In every case, the neighborhood method produced a lower error in occupancy estimates; it produced smaller errors in flow estimates in 10 of 15 locations. Overall, the neighborhood method performed better in the mean and median, as expected.

5.5 Real time detection

We have seen that loop errors are not distributed uniformly among all loops, and that most errors come from a few very bad loops. Furthermore, these loops remain bad for a long time. This is why the daily detection algorithm can remove most of the bad data while keeping most of the good data. There are, however, detectors that alternate between giving good data and giving bad data. It may be better to diagnose each sample of

these detectors. Results from previous research suggest that real time detection is difficult, because given 30-second measurement has a large variability, and most values cannot be ruled out. However, we developed an algorithm that uses information from the detector's historical diagnoses and neighbor detector measurements.

5.5.1 Setup

Let $\Delta_i(t)$ be the indicator that the ith detector is bad at t. We want to estimate $\Delta_1(t), \Delta_2(t), ...$ using measurements $\tilde{X}_1(t), \tilde{X}_2(t), \ldots$. Here, we use $\tilde{X}_i(t)$ to represent each of the measured quantities $\tilde{Q}_i(t)$ and $\tilde{K}_i(t)$. Both flow and occupancy are modeled the same way, only the model parameters are different. A diagnosis is made based on flow and occupancy independently, and the final diagnosis of the sample at time t is good only if it is good in both flow and occupancy. In the following discussion, we omit the time index t, because all operations are performed at the same sample time.

5.5.2 Maximum likelihood

The real time diagnostics algorithm considers the relationship between neighbor measurements. It decides whether a sample from a detector is bad by comparing it against values at its neighbors. Jaimyoung Kwon suggested a maximum likelihood solution to $\mathrm{P}(\Delta_1, \Delta_2, \ldots | \tilde{X}_1, \tilde{X}_2, \ldots)$. It requires the joint distribution of the measurement vector $(\tilde{X}_1, \tilde{X}_2, \ldots)^T$. Since this vector can have more than 100 dimensions, it's hard to estimate their joint distribution from historical data. Kwon used a Gibbs Sampler to solve the problem. This is a Monte Carlo technique that approximates the joint distribution of $(\tilde{X}_1, \tilde{X}_2, \ldots)^T$ by sampling from conditional distributions $\mathrm{P}(\tilde{X}_i | \tilde{X}_j)$. But because this

method requires many iterations to converge, it is not well suited for real time operation.

5.5.3 Marginal method

John Rice proposed to simply take the imputed value from the neighbors and compare it with the measured value at the current location. If the imputed value $\hat{X}_i$ is close to the measurement $\tilde{X}_i$, declare the measurement good; otherwise declare it bad. The imputed value is calculated from only the neighbors that are good according to historical diagnosis. Note that confounding is possible here. If loops i and j are both good according to historical diagnosis, but each accuse the other of being bad at the current sample time, we can't tell who's telling the truth.

We use an approach similar to Rice's, using only marginal probability distribution involving each neighbor pair. In essence, we find the probability that the ith loop is good, conditioned on each of its neighbors. Define $A_i \stackrel{\text{def}}{=} \{j : (i,j) \text{ are neighbors}\}$. Let

$$S_{ij} \stackrel{\text{def}}{=} \mathrm{P}(\Delta_i = 0 | \tilde{X}_i, \tilde{X}_j),\ \forall j \in A_i \tag{5.9}$$

be the *score* of i according to j. This definition avoids confounding because S_{ij} and S_{ji} unambiguously specifies the probabilities given the measurements at i and j. S_{ij} is a continuous random variable between 0 and 1, instead of a binary decision, so it can quantify the degree of agreement between two neighbors. The median of the neighbor scores is used as a final score of i based on all of its neighbors:

$$S_i \stackrel{\text{def}}{=} \underset{j \in A_j}{\text{median}}(S_{ij}) \tag{5.10}$$

The diagnosis of the $\tilde{X}_i$ is derived from S_i by thresholding with $s^* = 0.5$:

$$\hat{\Delta}_i(t) \stackrel{\text{def}}{=} 1(S_i(t) < s^*). \tag{5.11}$$

The time index t was reinserted in (5.11) to emphasize that this diagnosis is made at each sample time.

5.5.4 Linear model

In Section 5.4 on imputation, we showed the linear relationship between neighbor measurements of occupancy and flow. Here we use the same linear model, with the same parameters as before.

$$X_i = \beta_0(i,j) + \beta_1(i,j)X_j + \xi(i,j), \tag{5.12}$$

where $\xi(i,j)$ is zero-mean normal, with variance σ^2_{ij}. When both samples are good, the measurements are assumed to be equal to the actual values:

$$\Delta_i = 0,\ \Delta_j = 0 \Rightarrow \left\{ \begin{array}{l} \tilde{X}_i = X_i,\ \tilde{X}_j = X_j, \\ \\ \tilde{X}_i \sim \mathcal{N}(\beta_0(i,j) + \beta_1(i,j)\tilde{X}_j, \sigma^2_{ij}), \end{array} \right\} \tag{5.13}$$

i.e. $\tilde{X}_i$ is normal with mean $\hat{X}_{ij}$ and variance σ^2_{ij}, where

$$\hat{X}_{ij} = \beta_0(i,j) + \beta_1(i,j)\tilde{X}_j \tag{5.14}$$

is the imputed value from j. The parameters β_0, β_1 are the same as those estimated in the imputation section.

To solve for S_{ij}, we need the conditional density of $\tilde{X}$ when the loops are bad. When either $\Delta_i = 1$ or $\Delta_j = 1$, we model $\tilde{X}_i$ as independent of $\tilde{X}_j$. We assume that $\tilde{X}_i$ has a uniform density when $\Delta_i = 1$.

$$\mathrm{P}(\tilde{X}_i = x | \Delta_i = 1) = \left\{ \begin{array}{ll} \frac{1}{b-a} & \text{if } a \le x \le b \\ \\ 0 & \text{else} \end{array} \right. \tag{5.15}$$

When $\tilde{X}_i$ represents occupancy, the range of the uniform distribution is naturally $[0, 1]$; when $\tilde{X}_i$ is flow rate, we set the range to be $[0, 3000]$ vph, and declare any flow rate over 3000 vph to be bad. The uniform distribution is used because we assume nothing about the measurement of a bad detector. In the future, we may be able to improve the algorithm by modeling this distribution more accurately.

5.5.5 Neighbor scores

An expression for S_{ij} can be found by conditioning on the Δ's.

$$\begin{aligned} S_{ij} &= \mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_i, \tilde{X}_j)\mathrm{P}(\Delta_j = 0|\tilde{X}_i, \tilde{X}_j) + \\ &\quad \mathrm{P}(\Delta_i = 0|\Delta_j = 1, \tilde{X}_i, \tilde{X}_j)\mathrm{P}(\Delta_j = 1|\tilde{X}_i, \tilde{X}_j) \\ &= \mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_i, \tilde{X}_j)\mathrm{P}(\Delta_j = 0|\tilde{X}_i, \tilde{X}_j) + \\ &\quad \mathrm{P}(\Delta_i = 0|\tilde{X}_i)(1 - \mathrm{P}(\Delta_j = 0|\tilde{X}_i, \tilde{X}_j), \end{aligned} \tag{5.16}$$

the last step follows because $\tilde{X}_i, \tilde{X}_j$ are independent given $\Delta_i = 1$. Let

$$U_i \stackrel{\text{def}}{=} \mathrm{P}(\Delta_i = 0|\tilde{X}_i); \tag{5.17}$$

$$V_{ij} \stackrel{\text{def}}{=} \mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_i, \tilde{X}_j), \tag{5.18}$$

Equation (5.16) becomes

$$S_{ij} = V_{ij}S_{ji} + U_i(1 - S_{ji}). \tag{5.19}$$

Solving for S_{ij}, we get

$$S_{ij} = \frac{U_jV_{ij} - U_jU_i + U_i}{1 - V_{ji}V_{ij} + V_{ji}U_i + V_{ij}U_j - U_iU_j}. \tag{5.20}$$

Now we focus on the conditional probabilities (5.17) and (5.18). Using Bayes Rule,

$$\mathrm{P}(\Delta_i = 0|\tilde{X}_i = x) = \mathrm{P}(\Delta_i = 0)\frac{f_{\tilde{X}_i|\Delta_i}(x|\Delta_i = 0)}{f_{\tilde{X}_i}(x)}, \tag{5.21}$$

where $f_{\tilde{X}_i}(\cdot)$ is the probability density function of $\tilde{X}_i$, and $f_{\tilde{X}_i|\Delta_i}(\cdot)$ is the conditional density of $\tilde{X}_i$ given Δ_i. It's hard to estimate $f_{\tilde{X}_i|\Delta_i}$, because we don't have a training set with known Δ_i's. Instead, we approximate (5.21) with the unconditional probability, $\mathrm{P}(\Delta_i = 0)$. Although this requires a training set as well, we make use of the statistics from historical data. We need the proportion of historical samples that are bad, whereas the statistics S_1, S_2, S_3 from the daily detection algorithm record the number of samples that are, in a way, *suspected* to be bad. We estimate the prior probability of a sample being bad as the ratio of suspected bad samples to all samples – for all t,

$$\mathrm{P}(\Delta_i(d,t) = 1) \approx \frac{1}{S_0(i,d-1)}(S_1(i,d-1) + S_2(i,d-1) + S_3(i,d-1)), \tag{5.22}$$

where $S_0(i,d-1)$ is the number of samples received from the ith detector on the $(d-1)$th day. Because of the bimodal nature of these statistics exhibited in figures 5.4 - 5.7, most of the priors are close to either zero or one. For these cases, the resulting score S_{ij} is strongly influenced by the prior. The real time detection algorithm provides extra information for those loops whose statistics fall between the extreme values.

We solve for V_{ij} in (5.18). Our model says that given $\Delta_i = 0$, $\tilde{X}_i$ follows a normal distribution whose mean depends on $\tilde{X}_j$; otherwise, $\tilde{X}_i$ is independent of $\tilde{X}_j$ and is uniformly distributed. We write the expression for (5.18) as

$$\begin{aligned} V_{ij} &= \mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_i, \tilde{X}_j) \\ &= \frac{\mathrm{P}(\tilde{X}_i|\Delta_i = 0, \Delta_j = 0, \tilde{X}_j)\mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_j)}{\mathrm{P}(\tilde{X}_i|\Delta_j = 0, \tilde{X}_j)}. \end{aligned} \tag{5.23}$$

In the numerator, the conditional density of $\tilde{X}_i$ is Gaussian given $\Delta_i = \Delta_j = 0$, with mean $\hat{X}_{ij}$ and variance σ_{ij}^2, as is given by the model in (5.13); the distribution of Δ_i is independent of both Δ_j and $\tilde{X}_j$. Therefore,

$$\mathrm{P}(\tilde{X}_i|\Delta_i = 0, \Delta_j = 0, \tilde{X}_j)\mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_j) = \phi(\tilde{X}_i, \hat{x}_{ij}, \sigma_{ij}^2)\mathrm{P}(\Delta_i = 0),$$

where $\phi(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right)$ is the normal density. The denominator in (5.23) is equal to

$$\begin{aligned} & \mathrm{P}(\tilde{X}_i|\Delta_j = 0, \tilde{X}_j) \\ = \; & \mathrm{P}(\tilde{X}_i|\Delta_i = 0, \Delta_j = 0, \tilde{X}_j)\mathrm{P}(\Delta_i = 0|\Delta_j = 0, \tilde{X}_j) \\ & +\mathrm{P}(\tilde{X}_i|\Delta_i = 1, \Delta_j = 0, \tilde{X}_j)\mathrm{P}(\Delta_i = 1|\Delta_j = 0, \tilde{X}_j) \\ = \; & \phi(\tilde{X}_i, \hat{x}_{ij}, \sigma_{ij}^2)\mathrm{P}(\Delta_i = 0) + u(\tilde{X}_i)\mathrm{P}(\Delta_i = 1), \end{aligned} \tag{5.24}$$

where $u(x)$ is the uniform distribution with range $[a, b]$ as in (5.15). The full expression for V_{ij} is

$$V_{ij} = \frac{\phi(\tilde{X}_i, \hat{x}_{ij}, \sigma_{ij}^2)\mathrm{P}(\Delta_i = 0)}{\phi(\tilde{X}_i, \hat{x}_{ij}, \sigma_{ij}^2)\mathrm{P}(\Delta_i = 0) + u(\tilde{X}_i)\mathrm{P}(\Delta_i = 1)}. \tag{5.25}$$

Figure 5.16 shows results of the real time detection on several neighboring loops. Five-minute occupancy measurements are shown, and the samples declared to be bad are marked by squares. Loop 219 is clearly bad. The only other bad loop seems to be loop 218. This loop behaved normally most of the time, but near 10:00 and after 20:00, the occupancy rises to 0.8, while other loops in the neighborhood showed low occupancies at the same sample times. The algorithm correctly identified these points. On the other hand, some of the points in loops 210, 211, and 216 were flagged when they appear to be good. These false positives occurred because measurements between neighbors loops are not as

correlated during congestion as they are when traffic is lighter. However, even though some good samples points will be labeled as bad, the imputation fills them with reasonable values. It may be more important to eliminate the truly bad samples, even if we remove a few good samples in the process.

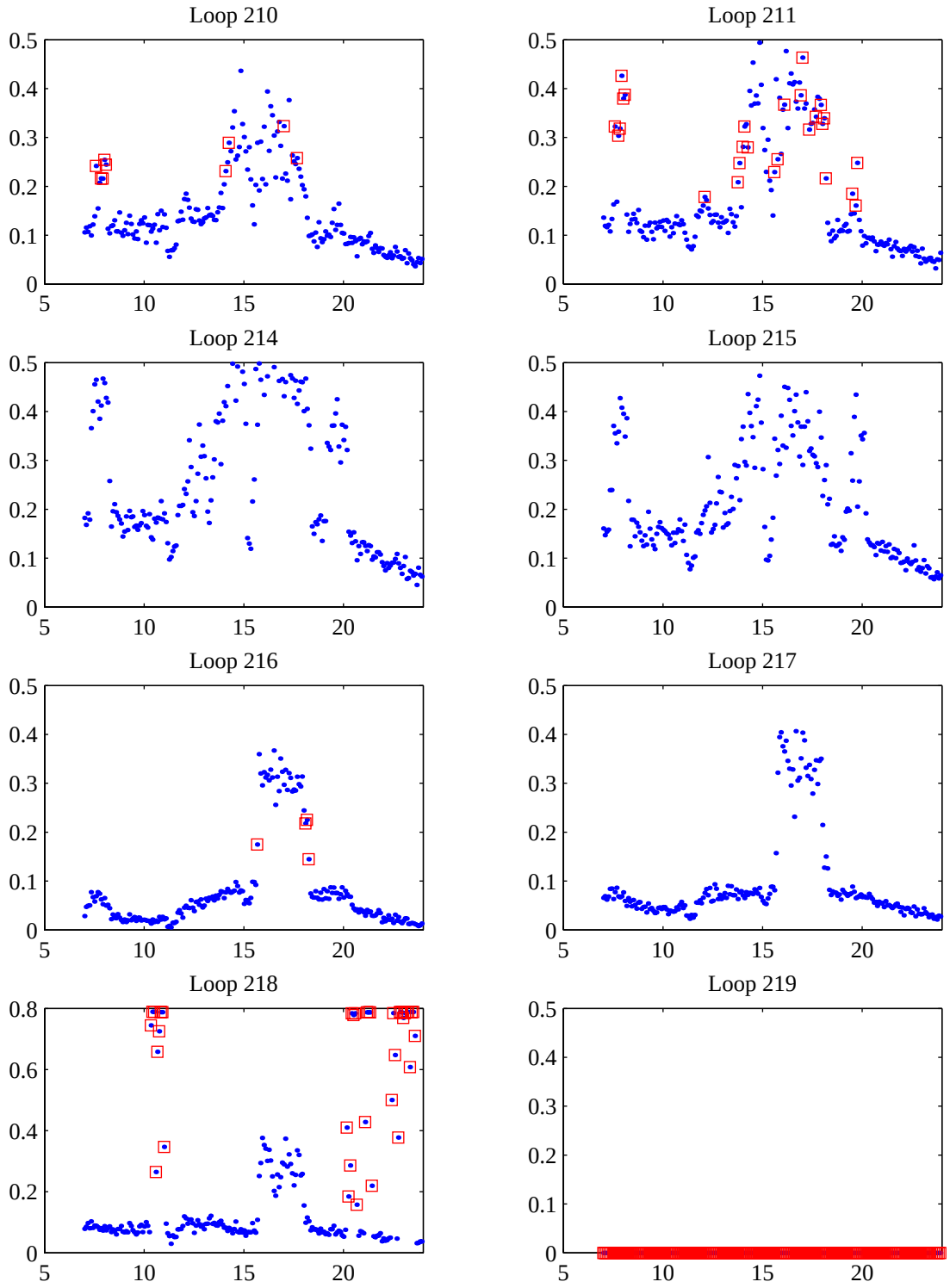


Figure 5.16: Real time diagnostics results.

Chapter 6

Estimate Speed From Single Loop Detectors

Speed is one of the most important quantities of traffic flow. We use it to calculate performance measures such as travel time and delay. Most loop detectors in California are single loops, which do not measure speed directly as double loops do. We need a way to estimate speed from single loop measurements of flow and occupancy. Any alternative method to measure speed directly requires the installation of other surveillance equipment, a daunting task that will take years to provide the coverage of existing single loop installations.

PeMS uses an algorithm that estimates the average length of vehicles at each location in real time, then estimates speed from the length. This algorithm was proposed by Ben Coifman and refined by Zhanfeng Jia, both of the PeMS development group. In comparisons with double loop measurements, this algorithm accurately estimates speeds and performs much better than existing methods. We call this the Jia-Coifman algorithm

and describe it in this section using the following definitions:

- $t = 0, 1, \ldots, 287$: five-minute sample number for each day;
- δ: length of the sample period in hours;
- $Q(t) \in \{0, 1, \ldots\}$: number of vehicles in the tth sample period;
- $K(t) \in [0, 1]$: average occupancy in the sample period.

Erik van Zwet developed an improved version of the algorithm which is presented at the end of this chapter.

6.1 Speed and length

The average vehicle speed in a sample period can be estimated from volume and occupancy if the average vehicle length is known. Let $L_i(t), V_i(t)$ be random variables representing the length and speed of the ith vehicle in the tth sample time. The relationship of flow, occupancy, length, and speed is

$$K(t) = \sum_{i=1}^{Q(t)} \frac{1}{\delta} \frac{L_i(t)}{V_i(t)}. \tag{6.1}$$

Let $\bar{L}(t)$ be the average length of vehicles $1, 2, \ldots, Q(t)$, and $\bar{V}(t)$ be their average speed, where

$$\bar{L}(t) = \frac{1}{Q(t)} \sum_{i=1}^{Q(t)} L_i(t) \tag{6.2}$$

$$\bar{V}(t) = \frac{1}{Q(t)} \sum_{i=1}^{Q(t)} V_i(t). \tag{6.3}$$

We want to estimate $\bar{V}(t)$. Let $\mu(t) \equiv \mathrm{E}[V_1(t)]$ be the expectation of the individual speeds. For large $Q(t)$, the law of large numbers says $\bar{V}(t) \approx \mu(t)$. Assuming that $V_i(t)$ and $L_i(t)$ are uncorrelated and iid in i, Dailey [39] showed that

$$\mu(t) \approx \frac{Q(t)\bar{L}(t)}{K(t)} \tag{6.4}$$

where $\sigma^2(t)$ is the variance of $V_1(t)$. In this relationship, Q and K are observed but $\bar{L}$ is unknown and need to be estimated.

6.2 Constant-length method

The traditional method of single loop speed estimation assumes a constant average vehicle length. For example, Caltrans uses $\bar{L}(t) = 18$ feet for all t and all loops. This implies that the vehicle mix at different locations and times are the same. However, the variation in $\bar{L}(t)$ is known to be large in [40]. Part of this variation comes from the ratio of trucks to passenger cars in the mix. For example, in the early morning period between 1:00 am and 5:00 am, there are very few passenger cars but relatively more trucks; during commuting hours, most of the vehicles are passenger cars. Because trucks are longer than cars, the average length of vehicles depends on their ratio.

Measurements from I-80 in Emeryville, California confirm that average length of vehicles in the same location varies as the time of day. Since there are double loop detectors at this location, $\bar{V}(t)$ is also measured along with $Q(t)$ and $K(t)$. Length is calculated using

$$\bar{L}(t) = \frac{\bar{V}(t)Q(t)}{K(t)}. \tag{6.5}$$

Figure 6.2 shows that $\bar{L}(t)$ varies from 21 feet and 40 feet during one day.

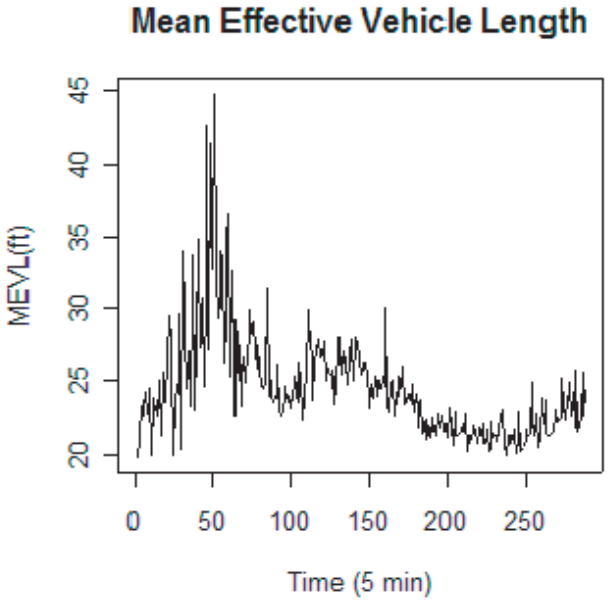


Figure 6.1: Average vehicle length on I-80 at one location over one day. We use "Mean Effective Vehicle Length" to emphasize that these are lengths derived from double loop speeds rather than directly measured.

Average vehicle lengths also vary by location – different routes have significantly different number of trucks on them. There is also lane-to-lane variation. In most locations, the left lane has almost no trucks and the right lanes have many more trucks. Yet another source of variation lies in the different sensitivities of different loops. The same vehicle mix can have different apparent average lengths when measured by different loops.

Let $\bar{L}(x,t)$ be the average vehicle length at location x and sample period t. We present the variation in $\bar{L}(x,t)$ in Los Angeles estimated from single loop measurements. Single loops are used instead of the double loops because we do not have enough double loops do perform a statistical study. Because lengths or speeds aren't directly measurable with single loops, lengths are estimated using the Jia-Coifman algorithm. The data are from 1340 lane-1 locations and 971 lane-3 locations. For each (x,t), we compute estimates

of estimate of $\bar{L}(x,t)$ which we call $\hat{L}(x,t)$. The average of $\hat{L}(x,t)$ in time and its variation from this mean are shown in Figure 6.2. These plots show the histograms of average length over one day for lanes 1 and 3, defined as

$$\bar{L}(x) = \frac{1}{288}\sum_{t=0}^{287}\hat{L}(x,t). \tag{6.6}$$

$\bar{L}(x)$ varies between 10 and 30 feet in both lanes 1 and 3, with medians of 14.4 feet for lane 1 and 16.9 feet for lane 3. Also, vehicle lengths in lane 3 are more variable than those in lane 1, because lane 1 contains mostly passenger vehicles all the time, and lane 3 usually has many more trucks, especially during early morning. during the day.

6.3 Adaptive algorithm

Because of the variation in $\bar{L}(x,t)$ for different locations and times, we cannot assume a constant average length for all times or all locations. We present an adaptive algorithm that tracks $\bar{L}(x,t)$ over time. This algorithm was called the "G-factor" algorithm in earlier publications, but because of confusion over whether the g is the vehicle length or its inverse, we call it the Jia-Coifman algorithm.

In the following analysis, we omit the location index x because the discussion applies to any location. The adaptive algorithm estimates $\bar{L}(t)$ when t is in free flow, and estimates speed during congestion. The key insight is the following free flow speed hypothesis.

Hypothesis 1 *Traffic at sample time t is either in free flow or in congestion. During free flow, speed is constant and equal to the free flow speed.*

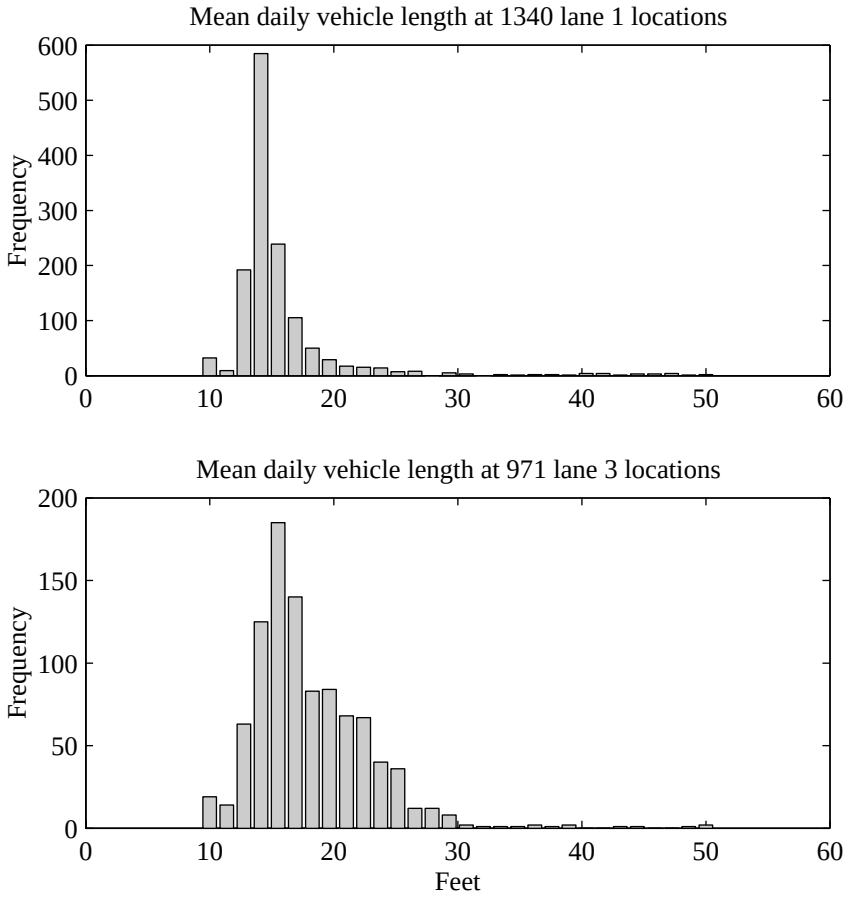


Figure 6.2: Variation in $\mathrm{E}[\bar{L}](t)$ in Los Angeles

Let $C(t)$ be the indicator that t is in congestion. The conjecture says that

$$\bar{V}(t) = v_f \text{ if } C(t) = 0. \tag{6.7}$$

Hypothesis 1 leads to the following equation to estimate length during free flow:

$$\hat{L}(t) = \frac{v_f \delta K(t)}{Q(t)} \text{ if } \hat{C}(t) = 0, \tag{6.8}$$

where $\hat{C}(t)$ is an estimate of the congestion state. We use a threshold on occupancy to

determine whether the traffic is in congestion,

$$\hat{C}(t) = \begin{cases} 1 & \text{if } K(t) > 0.15, \\ \\ 0 & \text{otherwise.} \end{cases} \tag{6.9}$$

During congestion, the speed is no longer v_f. If the average lengths of vehicles do not change quickly with time, then we can use the recent estimates of $\hat{L}(t)$ to approximate the current vehicle lengths, i.e.

$$\hat{L}(t) = \hat{L}(t-1) \text{ if } \hat{C}(t) = 1. \tag{6.10}$$

Speed is estimated from length by

$$\hat{V}(t) = \frac{\hat{L}(t)Q(t)}{\delta K(t)}. \tag{6.11}$$

Because $\hat{L}(t)$ is an instantaneous estimate based on only one sample time, the speed estimate in (6.11) has a large error. An exponential filter is used to smooth out the length estimates and produce

$$\hat{L}_{filt}(t) = \hat{L}(t)w + \hat{L}_{filt}(t-1)(1-w), 0 \leq w \leq 1. \tag{6.12}$$

The exponential filter is a simple way to implement real time filters because of its statelessness. As with any causal filter, it introduces a lag equal to the time constant of the filter. The lag τ is approximately

$$\tau = \frac{-1}{\log(1-w)}, \tag{6.13}$$

which is set to 1 hour in our algorithm. Because of the lag, the real lengths at time t is closer to $\hat{L}_{filt}(t+\tau)$. The difference $\hat{L}_{filt}(t+\tau) - \hat{L}_{filt}(t)$ is estimated using historical data. This is done as follows: compute $\hat{L}_{filt}(d,t)$, the filtered length estimates on the dth day and tth sample time, for $d = 1, 2, \ldots, n_d$; find the historical mean of the lengths for each time

of day,

$$\hat{L}_{hist}(t) \equiv \frac{1}{n_d} \sum_{d=1}^{n_d} \hat{L}_{filt}(d,t). \tag{6.14}$$

Approximate the time-advanced estimate as

$$\hat{L}_{filt}(t+\tau) \approx \hat{L}_{filt}(t) + \left[\hat{L}_{hist}(t+\tau) - \hat{L}_{hist}(t)\right]. \tag{6.15}$$

The Jia-Coifman algorithm is summarized below.

1. Estimate the instantaneous values of length for each sample, then filter it using an exponential filter:

$$\hat{L}_{inst}(t) = \frac{K(t)}{Q(t)} v_f \ \forall t \text{ s.t. } C(t) = 0 \tag{6.16}$$

$$\hat{L}_{filt}(t) = \begin{cases} (1-\exp(-\frac{1}{\tau}))\hat{L}_{inst}(t) + \exp(-\frac{1}{\tau})\hat{L}_{filt}(t-1) & \text{if C(t) = 0,} \\ \hat{L}_{filt}(t-1) & \text{otherwise;} \end{cases} \tag{6.17}$$

2. Correct for the lag using historical average of the length:

$$\hat{L}(t) = \hat{L}_{filt}(t) + [\hat{L}_{hist}(t) - \hat{L}_{hist}(t-\tau)], \tag{6.18}$$

where $\hat{L}_{hist}(t)$ is as in (6.14);

3. Estimate speed as

$$\hat{V}(t) = \hat{L}(t) \frac{Q(t)}{\delta K(t)}. \tag{6.19}$$

This algorithm was evaluated using double loop data from Interstate 80 in Emeryville, California, where there are direct measurements of speed as well as flow and occupancy. We also calculated the real average vehicle lengths $\bar{L}(t)$ using (6.5). Estimates of length and speed were calculated using (6.18) and (6.19) and were found to match measured values from double loops.

Figure 6.3 shows the real and estimated lengths at one location in lane 4 on one day. The dashed line represents measured $\bar{L}(t)$, the dotted line $\hat{L}_{filt}(t)$, and the solid line

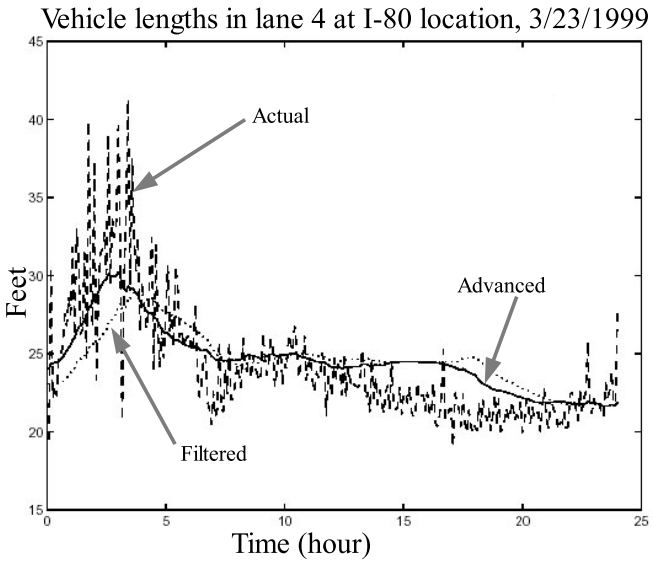


Figure 6.3: Instantaneous, filtered, and actual average lengths.

$\hat{L}(t)$, the time-advanced version. This plot shows that $\bar{L}(t)$ varies both in its trend and the variation around that trend. It also shows that $\hat{L}(t)$ follows the trend of the actual lengths well, whereas $\hat{L}_{filt}(t)$ lags by about an hour.

Figure 6.4 shows the estimated and actual speeds for one day. The measured speed is shown by the dashed line, and the solid line shows the estimated speed. The free flow speed is marked by the dotted line. This plot shows that the speed estimates are very close to measured values. They are almost perfectly superimposed on each other, especially during congestion. Figure 6.4 also exhibits the free flow phenomenon – speeds are almost constant in free flow, at 60 mph.

Also shown in Figure 6.4, however, there are unrealistically large variations in

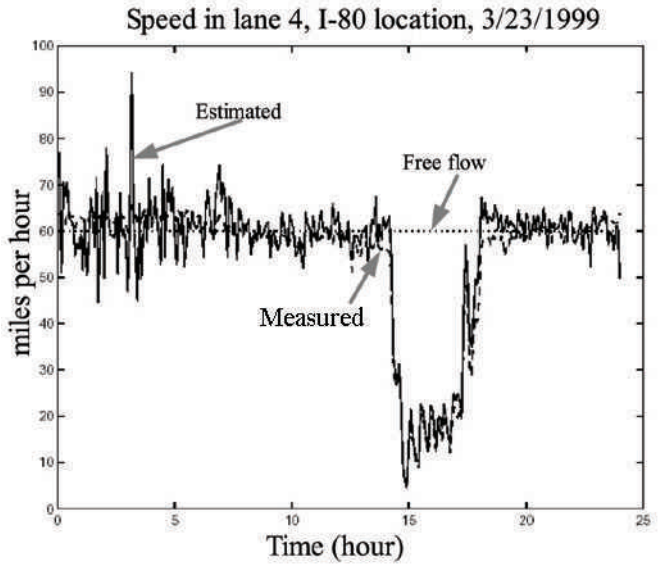


Figure 6.4: Speed estimated from average length

speed in the early morning. During this time, the freeway is in free flow and the actual speeds are close to 60 mph. Because of the low density of vehicles, the average length of vehicles varies greatly between sample periods, as shown in Figure 6.3. But because we use the filtered estimates of length, we cannot capture this variation in true length. The variation in Figure 6.3 shows up as variations in speed in Figure 6.4. This problem may be solved by filtering speed for several sample times.

We compare the performance of the adaptive algorithm with that of the single length algorithm. The comparison is done on double loop data from 20 loops on I-80. Estimates of speed and length are computed for each day $d = 1, 2, \ldots, n_d$ and location $x = 1, 2, \ldots, 20$. The constant-length estimates are calculated using

$$\hat{V}_{cl}(x, d, t) = l^*(x, d) \frac{Q(x, d, t)}{\delta K(x, d, t)}, \tag{6.20}$$

where

$$l^*(x,d) \equiv \frac{1}{288}\sum_{t=0}^{287}\hat{L}(x,d,t) \tag{6.21}$$

is the average length of all vehicles over the entire day d at location x. The Jia-Coifman algorithm is used to compute $\hat{L}(x,d,t)$ and $\hat{V}(x,d,t)$. To show the effect of the variation in length on speed estimates, we also computed the daily length variation $\hat{\sigma}_L(x,d)$, defined as the root mean squared difference of instantaneous length and the timed-mean length:

$$\hat{\sigma}_L(x,d) \equiv \sqrt{\frac{1}{288}\sum_{t=0}^{287}\left[\hat{L}(x,d,t) - l^*(x,d)\right]^2}. \tag{6.22}$$

The RMS error in speed estimates are plotted against $\hat{\sigma}_L(x,d)$ for each (x,d) in Figure 6.5. Each point is for one detector and one day. First, the average error of the Jia-Coifman algorithm is much lower than that of the constant-length method. Also, the error in the constant-length estimate grows linearly with the variation in the true lengths. This is expected because using a constant to estimate length will do poorly if the actual lengths are not constant. The adaptive algorithm is not very sensitive to $\hat{\sigma}_L(x,d)$ because it tracks changes in $\bar{L}(x,d,t)$.

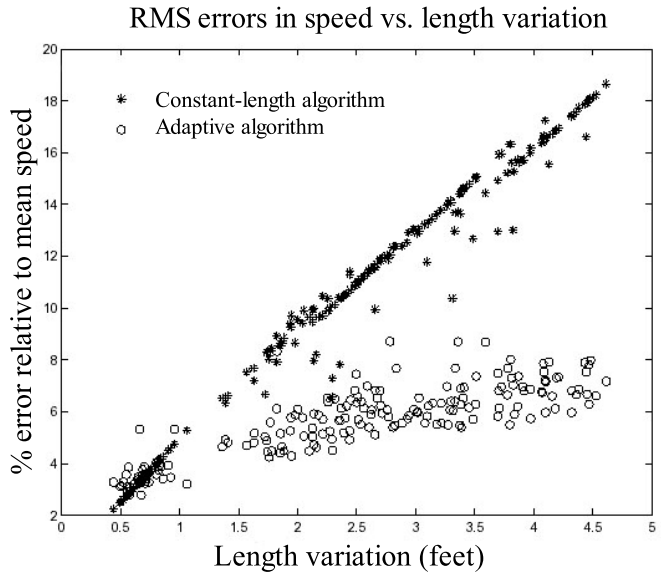


Figure 6.5: RMS error in speed estimates using the adaptive and constant length algorithms. Plotted against $\hat{\sigma}_L(x, d)$

6.4 Length profile algorithm

The Jia-Coifman algorithm is implemented and computes speeds for all the single loops in the PeMS database every five minutes. There are some problems with its results, however, and we have improved algorithm that tries to address some of its shortcomings.

6.4.1 Problems with adaptive algorithm

Although the adaptive speed algorithm performs well, it has a few problems. One of them was already illustrated in Figure 6.4. Because of filtering on the instantaneous length estimates, the resulting speed estimates have a large variation. Sometimes, speed estimates can reach 200 mph. This is because the average length in a 5-min period can vary greatly when the traffic volume is low.

Another problem involves the estimation of length during congestion. The adaptive algorithm currently takes $\hat{L}(t)$ from the most recent free flow period and uses this value throughout the congestion period. Again, for simplicity, we omit the location index x and day index d, with the understanding that the following discussion applies to any location and any day. It is crucial that we estimate $\hat{L}(t)$ correctly, because any mistake will be kept for the duration of the congestion period, which is exactly the most important period for speed estimation. The current algorithm uses a threshold on $K(t)$, set at 0.15, to flag congestion, see (6.9). However, we found that in different locations, congestion occurs at occupancies varying between 0.07 and 0.18. This variation is due to different traffic characteristics and detector sensitivities at different locations and lanes. In equation (6.17), the estimate depends on the correct diagnosis of congestion because of the assumption of

the free flow speed v_f. For example, suppose at a certain location, the freeway is already in congestion, and the traffic is flowing at 40 mph, but we think it is in free flow and 60 mph. The actual length is $\bar{L}(t) = 40\frac{Q(t)}{\delta K(t)}$, while the algorithm thinks it's $\hat{L}_{inst}(t) = 60\frac{Q(t)}{\delta K(t)}$. The error is

$$(60 - 40)/40 = 50\%.$$

Figure 6.6 shows the result of misdiagnosed congestion. Speeds for lanes 1, 3, and 4 are shown for this location. In lanes 1 and 3, the speed estimates look normal; in lane 4, the speed rises quickly to 80 mph just before congestion. This phenomenon occurs because of the overestimate of $\bar{L}(t)$ for the reasons mentioned above.

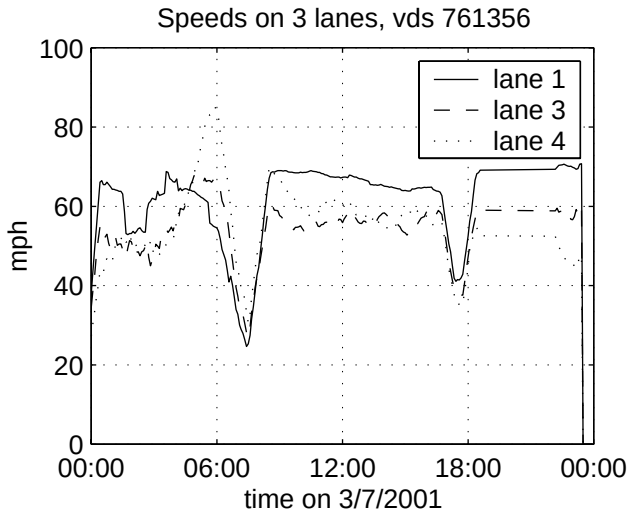


Figure 6.6: Bad speed estimate because of incorrect detection of congestion

6.4.2 Estimate daily length profile

An improvement to the adaptive algorithm was proposed by Erik van Zwet. We call this the Daily Length Profile Algorithm, or DLPA. This algorithm solves the two problems

noted above.

The DLPA starts with several assumptions. It models the vehicle length $L_i(x,d,t)$ as a random variable whose distribution depends only on the location x and the time of day t, but is iid in the day d. The basis for this assumption is that the same people drive on the same roads everyday, or at least on work days. Therefore, the distribution also depends on d, but only on whether it's a workday or a holiday. The DLPA finds the daily length profile of the average vehicle length for every time of day, and every type of days (weekday, weekend, holiday, day of the week, etc.), and use this profile as the estimate of the vehicle lengths in calculating the speed.

Let

$$v_f(x,t) \equiv \mathrm{E}[\bar{V}(x,d,t)|C(x,d,t)=0] \tag{6.23}$$

be the free flow speed at location x and time t. Here, the expectation is taken for each x and t – it's a function of (x,t) and doesn't depend on d. Assume $v_f(x,t)$ is known. In the implementation, we assume that free flow speeds depend only on the lane number not on time or postmile location. We also assume that $\bar{V}$ is uncorrelated with Q and K when in free flow. This is reasonable, since when there is no congestion, people's choice of speed doesn't depend on the density of vehicles. From (6.4), and fix the location x and the time of day t,

$$\begin{aligned}
\bar{L}(x,d,t) &= \bar{V}(x,d,t)\frac{\delta K(x,d,t)}{Q(x,d,t)}\ \forall x,d,t\ \Rightarrow \\
&\quad \mathrm{E}[\bar{L}(x,d,t)|C(x,d,t)=0] \\
&= \mathrm{E}\left[\bar{V}(x,d,t)\ \frac{\delta K(x,d,t)}{Q(x,d,t)}\middle| C(x,d,t)=0\right] \\
&= \mathrm{E}\left[\bar{V}(x,d,t)\,|C(x,d,t)=0\right]\mathrm{E}\left[\frac{\delta K(x,d,t)}{Q(x,d,t)}\middle| C(x,d,t)=0\right]
\end{aligned}$$

$$= \quad v_f \mathrm{E}\left[\frac{\delta K(x,d,t)}{Q(x,d,t)}\middle| C(x,d,t)=0\right]. \tag{6.24}$$

Equation (6.24) is the expectation of the average vehicle length at each time of day when there is free flow. Assuming the average vehicle length doesn't depend on $C(x,d,t)$, this is simply the expected length for each time of day, i.e.

$$\mathrm{E}\left[\bar{L}(x,d,t)\right] = \mathrm{E}\left[\bar{L}(x,d,t)\,|C(x,1,t)=0|\right]. \tag{6.25}$$

This is the daily length profile of vehicles at location x.

For each location x, we estimate (6.24) from observations at (d,t) where d's are the days and t's are the times when the data were collected. We have historical observations of $K(x,d,t)$ and $Q(x,d,t)$ at (d,t), as well as estimates of $C(x,d,t)$ in $\hat{C}(x,d,t)$. When $Q(x,d,t)>0$, we also have $v_f\delta\frac{K(x,d,t)}{Q(x,d,t)}$. These observations are plotted versus t in Figure 6.7. The LOESS method [41] is used to fit a smooth profile to this scatter plot. LOESS is a robust regression technique that produces estimates of the smooth function underlying the scatter plot. The LOESS estimate is shown in Figure 6.7 and estimates the trend of the scatter plot very well.

When estimating (6.24), we can be conservative in estimating $\hat{C}(x,d,t)$. The adaptive algorithm needs accurate estimates of $\hat{C}(x,d,t)$ because mis-diagnoses in both directions are costly. The DLPA needs to eliminate congestion periods, but doesn't care if some uncongested samples were thrown out as well, as long as there are enough samples to estimate $\mathrm{E}\left[\bar{L}\right]$. We choose the threshold for congestion to be the 60th percentile of all measured occupancy at each location, with the assumption that there is free flow at least 60% of the time. $\hat{C}(x,d,t)$ is estimated as in (6.9).

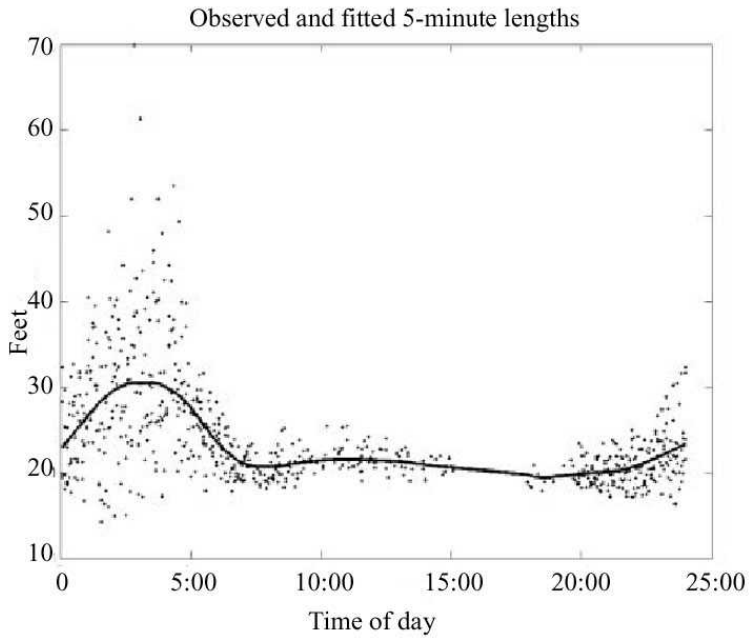


Figure 6.7: Scatter plot of $v_f \delta \frac{K(x,d,t)}{Q(x,d,t)}$ from five days, and the LOESS fit for time of day between 0 and 24 hours.

From the length profile, the speed estimate is

$$\hat{V}(x,d,t) = \frac{Q(x,d,t)}{\delta K(x,d,t)} \hat{L}(x,d,t). \tag{6.26}$$

The actual speed is

$$\bar{V}(x,d,t) \approx \frac{Q(x,d,t)}{\delta K(x,d,t)} \bar{L}(x,d,t), \tag{6.27}$$

and the error is

$$\begin{aligned} & \sqrt{\mathrm{E}[\bar{V}(x,d,t) - \hat{V}(x,d,t) | Q(x,d,t), K(x,d,t)]^2} \\ = \; & \frac{Q(x,d,t)}{\delta K(x,d,t)} \sqrt{\mathrm{E}[\bar{L}(x,d,t) - \hat{L}(x,d,t)]^2}. \end{aligned} \tag{6.28}$$

Because $\hat{L}(x,d,t)$ is close $\mathrm{E}[\bar{L}(x,d,t)]$, the error in speed is proportional to the standard deviation of the average vehicle length in a sample period. When $Q(x,d,t)$ is large, $\mathrm{E}[\bar{L}(x,d,t) - \bar{L}(x,d,t)]^2$ is small by the law of large numbers. When $Q(x,d,t)$ is small,

such as during the early morning periods, the length becomes more variable, as shown in Figure 6.3. We deal with this problem by filtering the estimated speed, such that if the current sample has very few vehicles, we estimate the speed to be a weighted sum of the previous estimate and the current estimate. The DLPA estimate of speed is

$$\hat{V}_{DLPA}(x,d,t) \stackrel{\text{def}}{=} w(Q(x,d,t))\hat{V}(x,d,t) + (1 - w(Q(x,d,t))\hat{V}(x,d,t-1). \tag{6.29}$$

In the above, the relative weight of the current estimate depends on the number of vehicles in the current sample period. The weight function chosen to be

$$w(q) = \begin{cases} 1 & \text{if } q > q^* \\ \frac{q}{q^*} & \text{else.} \end{cases} \tag{6.30}$$

This function assigns a weight of 1 to the current speed estimate if this estimate is based on a large enough sample size. We set $q^* = 1800$ vplph. The weight is between zero and 1 if $Q(x,d,t)$ is below q^*. Figure 6.8 shows the unfiltered speed estimate using (6.26) compared with the actual speeds. The estimate performs well in congestion, but is very noisy during hours when traffic is light. Figure 6.9 shows the filtered estimates using (6.29). The large oscillations are gone, and the estimate tracks the actual speeds almost exactly.

6.4.3 Discussion

The DLPA was tested on double loop data and was shown to perform better than the Jia-Coifman algorithm, and both are better than the constant-length algorithm. First, the filtering method of (6.30) removed the large variations in the speed estimates. Second, because we no longer need to detect the exact onset of congestion, we can be conservative with the threshold on occupancy when declaring congestion state.

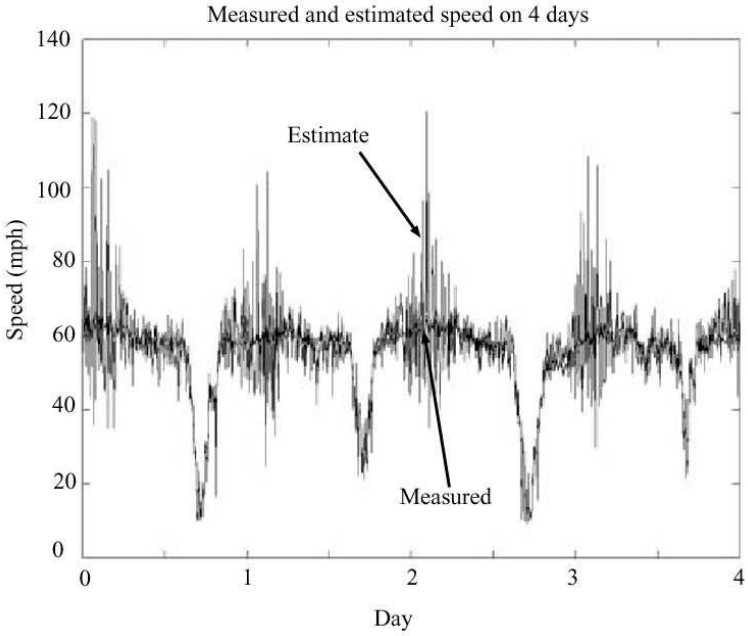


Figure 6.8: Speed estimate and actual speeds without filtering. Notice the large deviations during early morning.

One concern with this algorithm is its static nature. While it may work very well for most of the time, on days when the vehicle mix is different from usual, it could give the wrong results. But on the whole, this algorithm performs better than the alternatives.

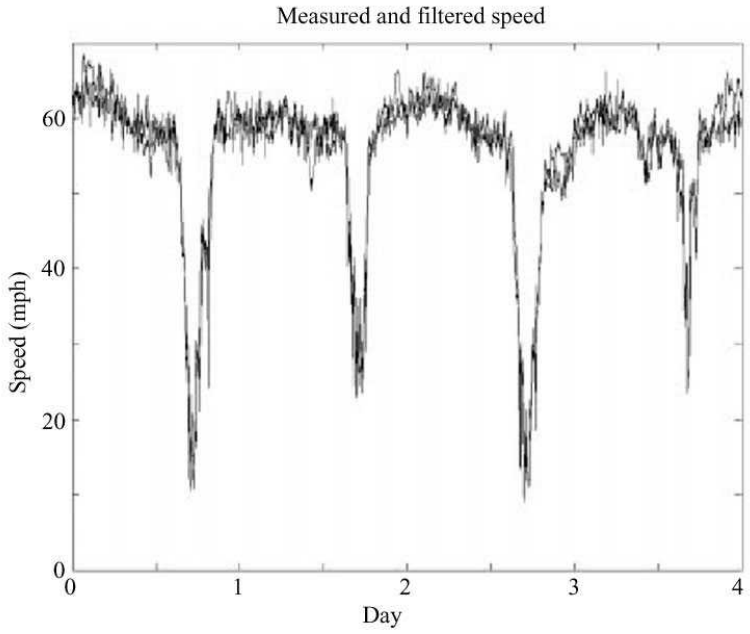


Figure 6.9: Speed estimate after filtering, compared to actual speeds.

Chapter 7

Travel Time Prediction And Routing

In Chapter 2 we described the ATIS prototype in PeMS. This service finds the quickest routes between two points based on predicted travel time. This is a classical shortest path problem. We model the freeway network as a connected graph with nodes and edges, where each node is the intersection of two freeways, and each edge is a freeway segment between two nearest nodes. Each edge e is assigned a weight $w(e,t)$ which is the time it takes to traverse the link departing at time t. Given a starting node o, a destination node d, and starting time t, we want to find the shortest paths between o and d, which minimize the total path weight. Our problems are 1) find the edge weights $w(e,t)$ for all edges e and all times t, where t is in the future, and 2) find the least-weight path given these path weights and the freeway network. These two parts are described in the next two sections. The work on travel time is based on work by Xiaoyan Zhang, Erik van Zwet, and

John Rice, while some of the experiment results are by the author. The shortest path section is based on work by the author with Professor Alistair Sinclair of the EECS department at UC Berkeley.

7.1 Travel time prediction

We have shown that travel time is very unpredictable for peak times (Section 3.2). Figure 7.1 shows the travel time profiles of 22 weekdays on a 40-mile corridor of Interstate 10 East. Each curve represents one day, and each point on the curve represents the travel time

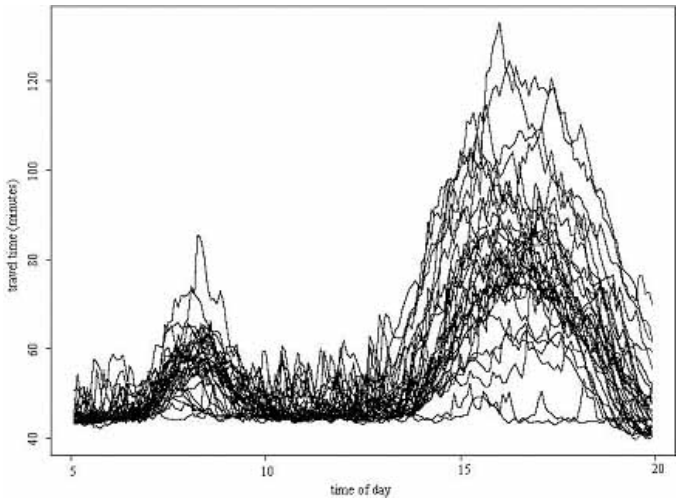


Figure 7.1: Travel time variation on Interstate 10, 40 miles

for a departure time. At 4 pm, the travel time varies between 42 minutes and 120 minutes! This is the variability if one knows only historical information. Can current information reduce the uncertainty? This picture suggests that it can. The days whose travel times are the highest at 4 pm also have high travel times at later times, which means that there are some temporal correlations between travel times on the same day. Traffic theory also

supports this idea - when a big backup occurs, its effect lingers for many hours even after the original cause of the congestion is removed. We find that using current loop measurements can significantly reduce the uncertainty in travel time.

7.1.1 Predict travel time on a segment

We would like to predict the travel time $T(t)$ on a segment for a trip departing at time t, where t is in the (near) future. Our measurements are speed estimates at detectors located on the segment at positions $x_1, x_2, \ldots, x_n$, for sample times t_i, $j = 1, 2, ...$ Therefore we collect V_{ij} at the ith time from the jth detector. Figure 7.2 shows a segment with n detectors. Let l_j be the length of the jth subsegment – the part of the segment that is

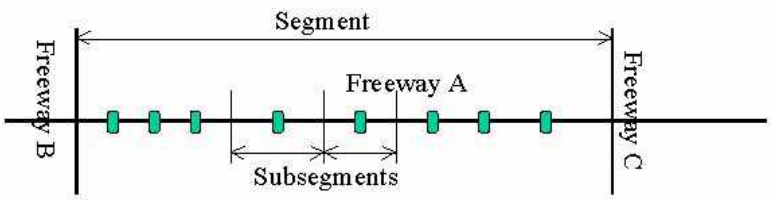


Figure 7.2: Segment and subsegments.

closest to x_j. At sample time t_i, we can approximate the travel time as

$$T^*(t_i) = \sum_{j=1} n \frac{l_i}{V_{ij}} \tag{7.1}$$

We call $T^*(s)$ the *instantaneous travel time* because it is based on instantaneous measurements. This would be the actual travel time if the speeds on the subsegments remain constant over the duration of the trip. In reality, speeds may change on the subsegments before the vehicle arrives there. To know the actual travel time, we need to know the future speed. While this is not possible in real time, we can estimate travel times in the past using

the method described in Section 3.2.2. We call this estimate $T(t)$, and use it as measured travel time. Section 3.2 showed that $T(t)$ is a close approximation of actual travel time. We found that $T^*(t)$ and $T(s)$ have a linear dependence, which is shown in Figure 7.3. This is

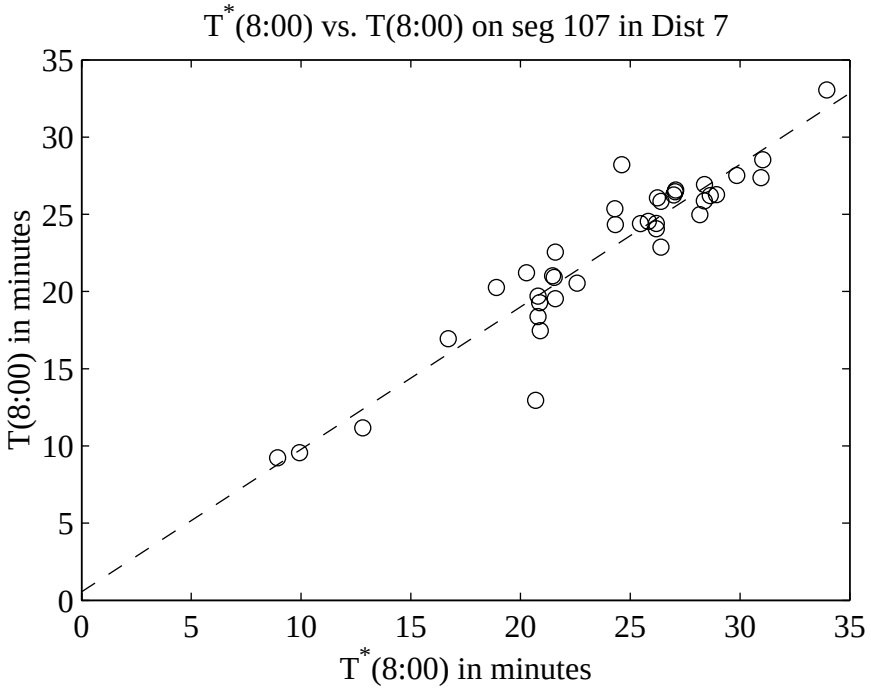


Figure 7.3: Linear relationship between T^* and T.

a scatter plot of $T(t)$ versus $T^*(t)$ on one segment in Los Angeles using data from August 2002. The departure time as well as the measurement time is 8:00 AM. This plot is very linear, and the linearity is strong long as $t - s$ is small. This suggests that linear regression is a good way to predict T using T^*. We model $T(t)$ as

$$T(t) = \beta_0(t_0, t - t_0) + \beta_1(t_0, t - t_0)T^*(t_0) + \epsilon(t) \tag{7.2}$$

where $\epsilon(t)$ is a zero-mean, iid random variable. The parameters β_0, β_1 depend on the current sample time t_0 and the *lag time* $t - t_0$. For example, if at 3 pm we want to predict the

travel time for a departure time of 4 pm of the same day. Then $t_0 = 3\text{pm}$, $t = 4\text{pm}$, and the lag is 1 hour. Intuitively, $\boldsymbol{\beta} \equiv (\beta_0, \beta_1)^T$ depends on the lag. We expect the travel time to be heavily dependent on T^* when the lag is small, but the current condition is not very predictive when the lag is large. This dependency also changes with the time of day because of very different traffic characteristics over the day. From the historical data, we can compute historical values of T and T^*, and use them to fit the parameters β_0, β_1. We found $T(d,t)$ and $T^*(d,t)$ for 22 weekdays on this segment, where $d = 1, 2, \ldots, n_d$ and $t = 0, 1, \ldots, 287$ are the sample numbers of 288 5-minute sample periods in a day. We estimated $\boldsymbol{\beta}$ using the least squares method. For each $\delta = 0, 1, 2, \ldots$ and each $t = 0, 1, 2, \ldots$, estimate

$$\hat{\boldsymbol{\beta}}(t,\delta) = \underset{\beta_0,\beta_1}{\arg\min} \sum_{1\le d\le n_d} |T(d, t+\delta) - \beta_0 - \beta_1 T^*(t)|^2 \tag{7.3}$$

Our implementation uses weighted least squares to impose smoothness on $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \underset{\beta_0,\beta_1}{\arg\min} \sum_{1\le d\le n_d, 0\le s\le 287} |T(d,s) - \beta_0 - \beta_1 T^*(t)|^2 \, g(t+\delta-s) \tag{7.4}$$

where $g(t)$ is a Gaussian kernel with mean zero and standard deviation set to 10 minutes. The weight function forces $\hat{\boldsymbol{\beta}}(t,\delta)$ to be smooth in t and δ. This condition is imposed because physically, we expect traffic behavior to change gradually with time.

The estimated parameters β_0, β_1 are used to predict travel time in real time, using

$$\hat{T}(t) = \hat{\beta}_0(t_0, t-t_0) + \hat{\beta}_1(t_1, t-t_0) T^*(t_0) \tag{7.5}$$

For example, if at 3 pm, we want to predict the travel time departing at 4 pm, then $t = 4\text{pm}$, $t_0 = 3\text{pm}$, and $\delta = 1\text{hour}$.

We applied (7.5) to the training data from I-10W, and compared its performance against those of historical mean travel time and $T^*(t)$. Let $\bar{T}(t)$ be the historical mean

travel time departing at t of a given day, where

$$\bar{T}(t) = \frac{1}{n_d} \sum_{1 \le d \le n_d} T(d,t), \ \forall t.$$

This is a reasonable prediction of the travel time on the current day if no real time information is available. $T^*(t)$ is also a natural estimator of the current travel time. This estimator uses only current information but no historical information. Using the historical data, we computed the root mean squared (RMS) error of these three estimators $\hat{T}(t)$, $\bar{T}(t)$, and $T^*(t)$. The results for the zero-lag and 60-minute lag cases are shown in Figures 7.4 and 7.5. In both cases, the prediction error is greatest for all schemes when departure time is during the afternoon peak hours. The historical mean estimator is not affected by the lag because it uses no real time information. These figures show that $\hat{T}(t)$ performs better than the other two schemes for both values of lag, for all departure times. The performance advantage is most apparent when lag is 60 minutes, when the maximum RMS error of the regression scheme is about 10 minutes, but the maximum error of the other schemes are about 18 and 25 minutes.

7.1.2 Alternative prediction methods

There are some alternatives to linear regression in predicting travel times. One of them is multiple regression using principal components. Our algorithm uses $T^*(t_0)$ as the predictor. This quantity represents a snap shot of the freeway at the most recent sample time t_0. It's reasonable to believe that the previous measurements also contain information about the future. We can model $T(t)$ as a linear combination of $T^*(t')$, $t' = t-0, t-1, t-2, \ldots$. This method captures more information about the system than the

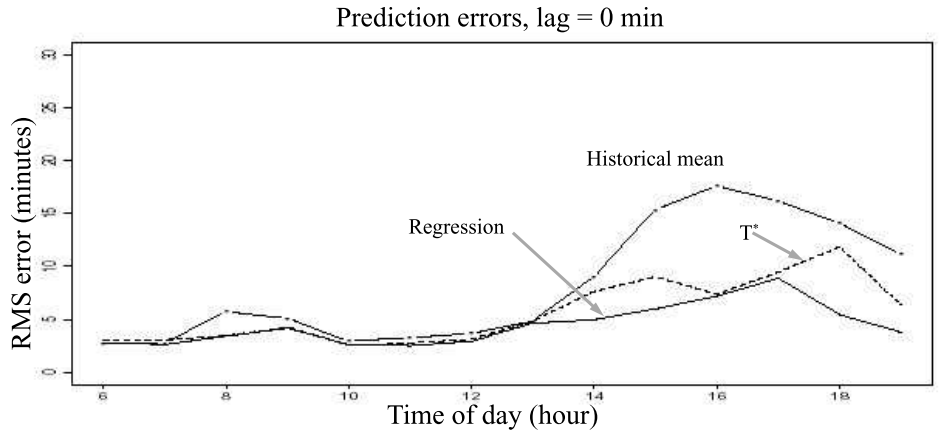


Figure 7.4: Prediction errors of regression, historical mean, and current status methods at lag = 0. Data from I-10W on 22 weekdays.

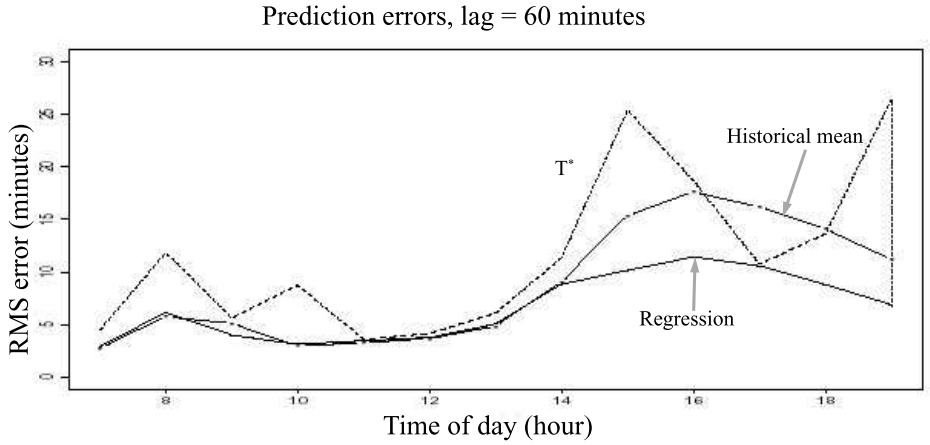


Figure 7.5: Prediction errors at lag = 60 minutes.

linear regression using only the current measurement. The principal components method reduces the number of parameters by approximating the covariance matrix of $T(t), T^*(t')$ with its largest eigenvectors.

Another method of prediction is the nearest neighbor method. In this method, one measures the trajectory of $T(d_0, t)$ or $T^*(d_0, t)$ of the d_0, the current day, up until the

current time t_0. When a prediction is needed for $T(t), t > t_0$, we find in all the historical trajectories of $T(d,t)$ or $T^*(d,t)$ the one that is "closest" to the current trajectory. We can define a measure of closeness between two trajectories, for example by the squared difference between them. Let d^* be the day with the closest trajectory to the current day, then we use $T(d^*,t)$ to predict $T(d_0,t)$. We implemented both the PCA and nearest neighbor methods and found the linear regression worked better than both of them. In the case of nearest neighbors, we may not have enough training data to obtain good results; in PCA, we may be over-fitting the data with too many parameters. In any case, the linear regression method works very well already, and significantly reduces uncertainty in travel time.

This method can be used to predict future travel times on any segment. Recall that we defined the Los Angeles freeway network as a connected graph with edges and vertices, where every edge is a freeway segment. To find the shortest route between two vertices, we need the edge weights $w(e,t)$. If we want to find the quickest route in travel time, we need to use travel time as edge weights:

$$w(e,t_0,t) = \hat{T}(e,t_0,t) \tag{7.6}$$

where $\hat{T}(e,t_0,t)$ is the predicted travel time for departure time t, and most recent sample time t_0, on edge e. $\hat{T}(e,t_0,t)$ is computed as in (7.5):

$$\hat{T}(e,t_0,t) = \hat{\beta}_0(e,t_0,t-t_0) + \hat{\beta}_1(e,t_0,t-t_0)T^*(e,t_0) \tag{7.7}$$

where $\hat{\boldsymbol{\beta}}(e,t,\delta)$ are the estimated model parameters for each segment e. These parameters for all segments are computed off line and stored in database tables.

7.1.3 Prediction on multiple segments

In a routing application, we need to predict the travel times on routes between any two vertices in a connected graph. Because there are many paths in a transportation network, we cannot individually model their travel time behavior using the linear model. Furthermore, for each origin-destination pair, we need to predict travel times on multiple routes. The only practical way to predict travel times on all routes is to build up from segment predictions. To do this, we have to combine predicted travel times on the segments that make up the graph. Suppose a path is made up of segments $1, 2, \ldots, n$, and we want to predict the travel time departing at t with current time t_0. Let $\hat{D}(e)$ be the predicted departure time of the eth segment, and $\hat{D}(1) = t$. The departure times of each segment e is given by

$$\hat{D}(e+1) = \hat{D}(e) + \hat{T}(e, t_0, \hat{D}(e)).$$

The path travel time is $\hat{D}(n+1)$.

We present the performance of this algorithm on 9 routes in Los Angeles using historical data. These routes are between 13.5 and 28.2 miles long. Information about the test routes is shown in Table 7.1. On each route, we compute the travel time at 15-minute intervals for a one-month period. We also compute the predicted travel times with prediction lags between 0 and 120 minutes. The root-mean-squared error between predicted travel time and actual travel times are presented.

We need to note that the parameters of prediction are calculated using the same data. Both prediction and model fitting used data from 23 weekdays from July 2002.

The prediction performed well using the iterative segment prediction. Figure 7.6

Route	Segments	Length (miles)	Average peak hour travel time (minutes)
1	3	23.7	47
2	3	28.2	40
3	3	28.2	46
4	4	14.9	26
5	4	14.9	23
6	4	13.6	19
7	4	13.6	23
8	3	16.4	33
9	3	16.4	36

Table 7.1: Route descriptions

shows the RMS error for various lags. On this route, which has 3 segments, the peak travel time occurs at 8 am. It has an average travel time of 50 minutes. The solid curve shows the prediction error if one has only knowledge of historical travel times. The dashed curve shows the prediction error of our linear regression method at a lag of zero. It reduces the uncertainty from 12 minutes to 4 minutes. As lag grows, the prediction becomes less precise. However, even at a lag of 60 minutes, the prediction still does better than historical mean.

We performed the same analysis on each of the 9 test routes. By looking at the performance of the prediction algorithm at various lags, we can find out how long the current information remains relevant. When the lag becomes too long, of course, the current measurement has very little predictive value, and the historical information becomes much more relevant. We summarized the performance of prediction on these routes using the RMS error during the peak hour. This is defined as the contiguous hour of day with the maximum mean travel time. The peak hour is used because the travel time is most variable when there is heavy congestion. Therefore, this is the period when prediction is

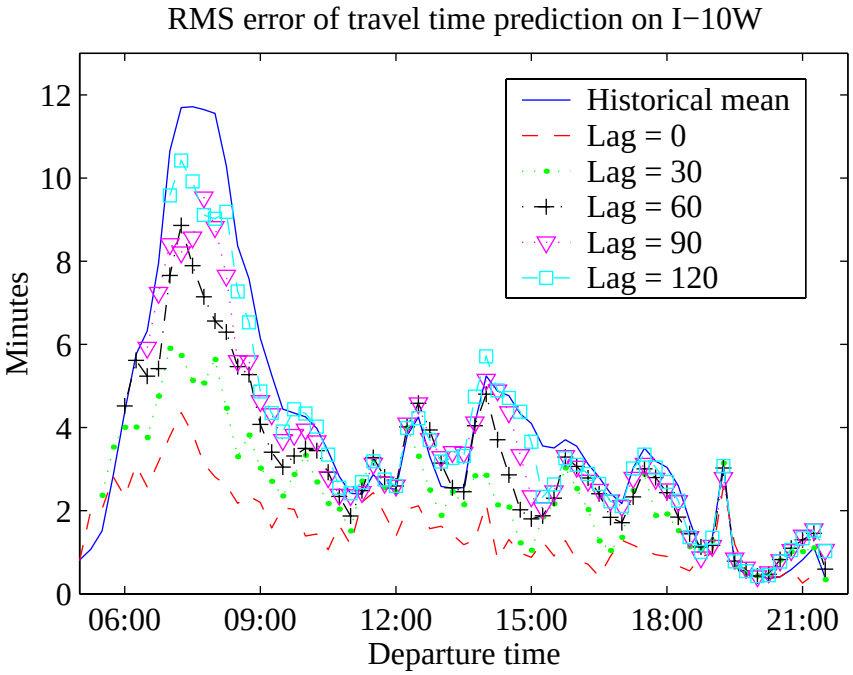


Figure 7.6: Prediction error at various lags, versus historical mean.

most valuable.

Figure 7.7 shows the median RMS error on these routes during their peak hours. The dashed line at 22% is the error using historical mean. This plot shows that the prediction becomes less accurate as the lag grows, as we expect. But at even very long lags, prediction performs better than using only historical average. For example, when predicting 30 minutes in advance, the error is less than 14% of the total travel time.

Figure 7.8 is the scatter plot of the prediction error at zero lag versus the travel time of each route. The prediction error grows linearly with average travel time.

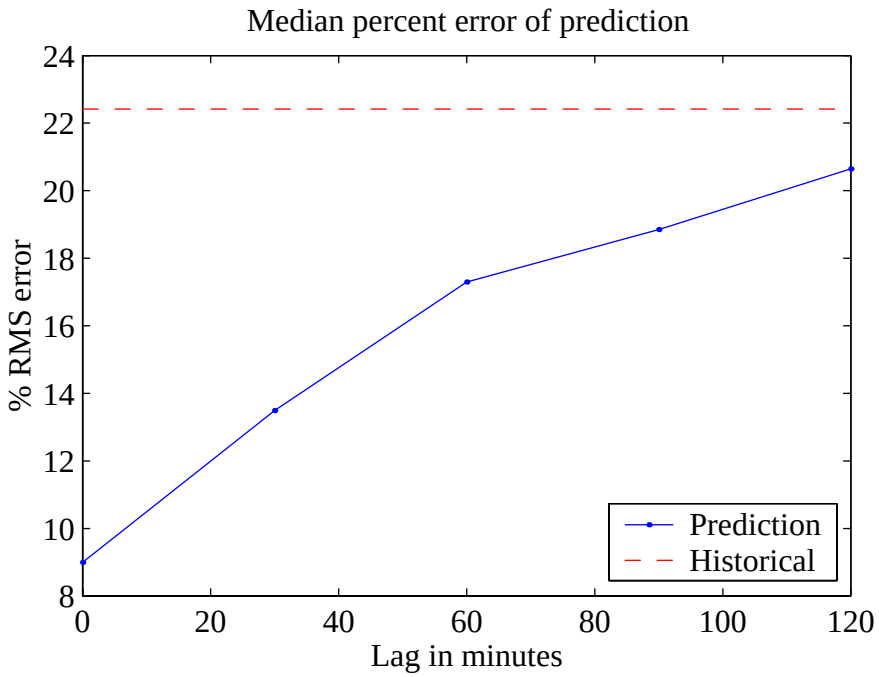


Figure 7.7: Median of prediction errors at various lags, versus historical mean.

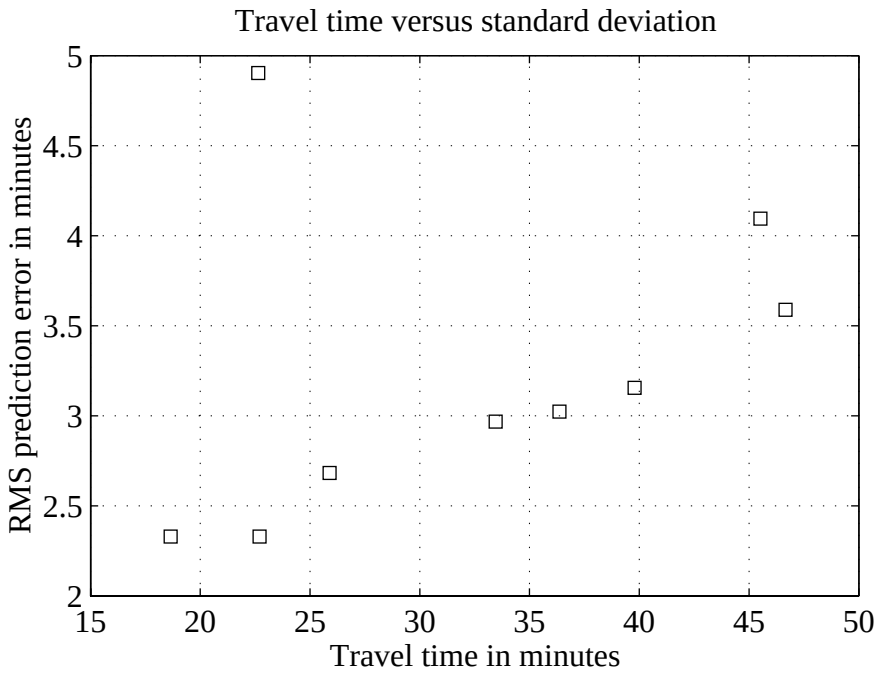


Figure 7.8: Prediction error plotted against average peak travel time.

7.1.4 Seasonal trends in model parameters

These results show the promise of travel time prediction using real time measurements. However, we may need to use more data to fit the parameters. The parameters used for this test were trained on one month of weekday data. However, a separate set of parameters is obtained for trips starting at each time of day. Therefore, even with the kernel smoothing, there are few samples to do the linear regression. Another potential problem is the seasonal shifts in traffic behavior. As we have noted in previous chapters, traffic tends to be better in the summer and heavier in the fall, at least in California. Therefore, parameters trained on data from July may not be valid for September. We may need to compute coefficients for different times of year, as well as time of day.

7.2 K shortest routes

The ability to predict travel times on any segment and route allows us to find shortest paths in the freeway network using the travel times as path weights. Often we want to know not only the quickest route, but several alternatives. The route guidance service in PeMS provides a list of routes and rank them in order of increasing travel time. The reason several alternative routes are provided rather than only the shortest one is because people may use other criteria than travel time. Plus, psychologically, people can see that the computer considered other routes, and that the recommended one is indeed the best among the candidates.

The implementation of this technology is straightforward. As the user enters the origin-destination pair o,d, and desired departure time, the computer finds the K best

routes between these points in the network using predicted travel times as path weights. Our implementation uses the segment lengths instead of travel time to first find a set of possible routes between each pair of vertices. The candidate routes between every pair of vertices are computed offline to speed real time operation.

The problem of finding K shortest paths, $K \geq 1$, between two vertices in a graph has been extensively studied. Specifically, we want paths that don't contain cycles, i.e. each edge in the path is traversed no more than once. Also, the edge weights in our network are always positive.

We implemented an algorithm that was presented by Yen [42]. Given a graph $G(V, E)$ with vertices V and edges E, this algorithm runs in $O(Kn^2 \log n)$ time, where $n \equiv |E|$ is the number of edges. We give a description of it here, and show how it's implemented.

We formally state the problem. Given a connected graph $G(V, E)$, with edge weights $w_1, \ldots, w_n$, where $0 \leq w_i \leq \infty$. Use $\boldsymbol{w} \stackrel{\text{def}}{=} (w_1, \ldots, w_n)$ to denote the vector of edge weights. Define a path by $\boldsymbol{x}$, such that $\boldsymbol{x} \stackrel{\text{def}}{=} (x_1, \ldots, x_n)^T$, where $x_i \in \{0, 1\}$ is the indicator that the path includes edge i. The path weight is defined as the sum of the edge weights:

$$f(\boldsymbol{x}) \stackrel{\text{def}}{=} \boldsymbol{w}'\boldsymbol{x} = \sum_{i=1}^{n} w_i x_i.$$

Let P be the set of all paths in $G(V, E)$ from o to d. Finding the shortest path can be posed as a constrained optimization, see Lawler [43]:

$$\text{minimize } f(\boldsymbol{x}), \text{ subject to } \boldsymbol{x} \in P.$$

Finding the K shortest paths, then, is equivalent to finding $(\boldsymbol{x}^{*1}, \ldots, \boldsymbol{x}^{*K})$ that minimizes

$$\sum_{k=1}^{K} f(\boldsymbol{x}^k),$$

subject to

$$\boldsymbol{x}^1 \neq \boldsymbol{x}^2 \neq \ldots \neq \boldsymbol{x}^K,$$

$$\boldsymbol{x}^k \in P \; \forall \, 1 \leq k \leq K.$$

Lawler and Yen showed how to find $\boldsymbol{x}^{*k}$ by successively partitioning P and removing the already found shortest paths. This procedure begins by finding the first shortest path, $\boldsymbol{x}^{*1}$. We used Dijkstra's algorithm in our implementation to find single shortest paths, which runs in $O(n \log n)$ time.

Using the following notation

$$\min(S) \equiv \arg\min_{\boldsymbol{x} \in S} f(\boldsymbol{x}),$$

the second shortest path is

$$\boldsymbol{x}^{*2} = \min(P \setminus \{\boldsymbol{x}^{*1}\}).$$

Suppose that, without loss of generality, $\boldsymbol{x}^{*1} = (1, 1, \ldots, 1, 0, \ldots, 0)$, i.e.

$$x_i^{*1} = \begin{cases} 0 & \text{for } 1 \leq i \leq h_1 \\ \\ 1 & \text{for } i > h_1 \end{cases},$$

where $h_1 = \sum_{i=1}^{n} x_i$ is the number of edges in $\boldsymbol{x}^{*1}$. We partition $P \setminus \{\boldsymbol{x}^{*1}\}$ by $P_1^1, P_2^1, \ldots, P_{h_1}^1$, where

$$P_j^1 \equiv \{\boldsymbol{x} : x_1 = 1, \ldots, x_{j-1} = 1, x_j = 0\} \cap P \tag{7.8}$$

It's clear that

$$P \setminus \{\boldsymbol{x}^{*1}\} = \bigcup_{j=1}^{h_1} P_j^1 \tag{7.9}$$

The meaning of the partitions is that the jth partition P_j^1 contains all paths from o to d that pass through edges $1, \ldots, j-1$, but do not pass through edge j. Figure 7.9 shows

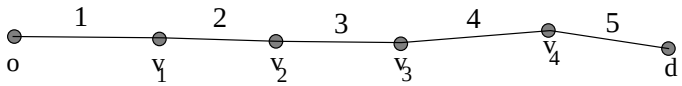


Figure 7.9: A sample path with 5 edges.

an example of the path, $\boldsymbol{x}^{*1}$. This path has 5 edges, so $h_1 = 5$. Take the 3rd partition, P_3^1. It contains all paths that use edges 1 and 2, but not 3. We can find the shortest path in this partition, $\min(P_3^1)$, by first finding the shortest path between v_2 and d using Dijkstra's algorithm, and adding to the beginning edges 1 and 2. For any $j \leq h_1$, we can find the minimum weight path in partition P_j^1 this way. The second shortest path is then

$$\boldsymbol{x}^{*2} = \min_{1 \leq j \leq h_1}(\min(P_j^1)). \tag{7.10}$$

This procedure is repeated until K shortest paths are found. At step k, the $k+1$st shortest path is found by further partitioning the partition that contains the kth shortest path. This is best illustrated graphically. In Figure 7.10, each node of the tree is a partition of P. At each step, the next shortest path is the shortest path of all the leaves of the tree. The partition that contains this path is further partitioned. Yen showed that paths found this way are the K shortest. This algorithm is efficient – each additional path requires no more than n Dijkstra's shortest path calculations. Therefore, the running time of the K shortest path algorithm is $O(Kn^2log(n))$.

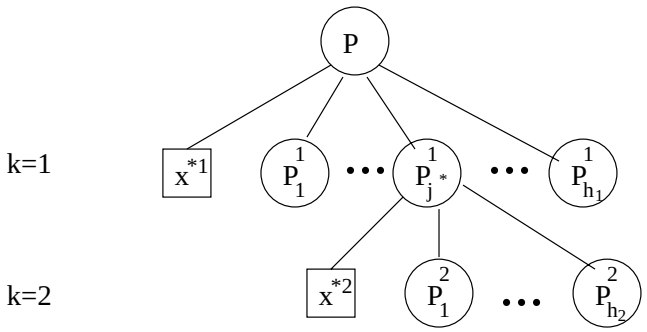


Figure 7.10: Procedure for finding K shortest paths, by successively partitioning the space of paths.

7.3 Implementation

To find the quickest route between a pair of vertices, we need to run the K shortest route algorithm in real time, using current measurements to predict future edge weights. This requires potentially many shortest path computations for each nth shortest path. On the other hand, we observe there usually aren't too many alternative freeway routes between two places. Therefore, we may be able to compute all the realistic alternative paths between every origin-destination pair in our network. Doing so simplifies the implementation and makes it faster. we construct N routes for each pair of vertices offline, with the expectation that the quickest routes at any time will be contained in this set.

The K shortest path algorithm was implemented in Perl on the network in Los Angeles. There are 41 vertices in this network, and between each pair of vertices, we found the 20 shortest paths based on driving distance. Therefore, we computed and stored $41^2 * 20 = 33,620$ paths in the database. This operation took about 20 hours to complete. Now that we have these static routes, the real time prediction and routing is easy. When a user requests the quickest route between any two points, we load the 20 paths from the

database corresponding to these origin and destinations, and predict the travel time on each route as described in Section 7.1.3.

The real time application is implemented as a Java Servlet. A servlet is a program that remains in memory and serves requests from the web server. Each time a user requests the routing service, the request first goes to the web server, which forwards it to the servlet. It is important for the servlet to stay in memory after serving a request because of the large overhead involved loading from the database the coefficients and current measurements. Therefore, the coefficients remain in memory and are refreshed periodically. Every 5 minutes, the application re-loads the parameters $\hat{\boldsymbol{\beta}}(e, t, \delta)$ for all segments e, the current measurement time t, and all lag times δ. It also loads from the real time data table

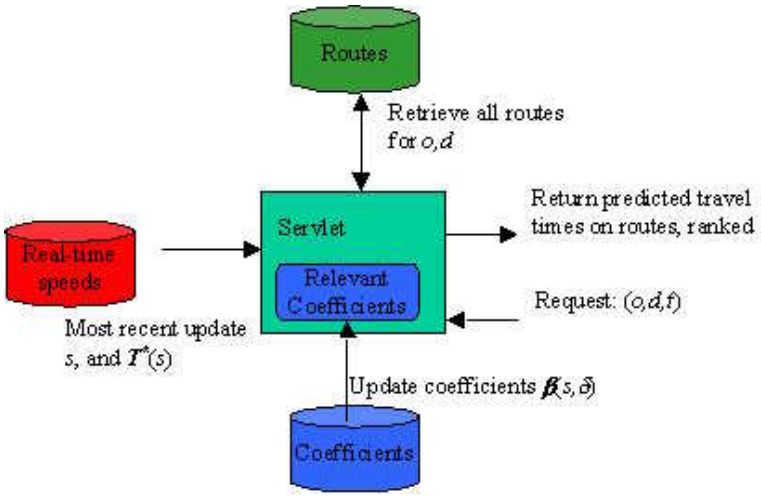


Figure 7.11: Servlet for travel time prediction and routing.

the current speeds to compute $T^*(e, t)$ on every segment for the current time. Each request requires no additional database actions, therefore speeding up the response. On our server, the response time for a quickest route computation is virtually instantaneous.

Currently, this service is available on the web to a computer based browser. But it is perhaps more useful if accessed on a mobile device. The low bandwidth requirement makes it a perfect application for palm top devices. Because the user only has to enter three pieces of information: origin, destination, and time of travel, and the result of the query are simply the route and the estimated time, the number of bits that need to be sent is very small. Today's 14 kbps wireless data channels are adequate to support this application. The more difficult task is in creating a good user interface. One way to simplify the data entry process is to have users define profiles at their leisure from a computer browser. For example, one can input his favorite routes, such as between home and work, to the mall, to the theater, etc. When he's in his car and in traffic, he needs only to select from a pre defined list of $o - d$ pairs to query alternate routes and travel times. These capabilities are described in [17]. The most appealing application of this technology is in telematic devices. Imagine stepping into a vehicle and have it ask you, "Where do you want to go?" and tells you the best route and that you will arrive at the destination in 24 minutes.

Advanced traveler information systems (ATIS) are an important application of ITS technologies. Knowing the best route saves time; knowing the travel time helps the driver make informed choices about his travel. As we demonstrated here, PeMS make it possible to provide traveler information.

Chapter 8

Conclusion

PeMS is an on-going project. It began in 1998, and was envisioned as an on-line tool for traffic management using real time and historical data. Today, it is fulfilling the promise of intelligent freeway operations using information technology.

Traffic congestion is an urgent problem in urban areas. While the number of vehicle miles driven rises every year, it's no longer possible to increase capacity by building more freeways in most cities. The answer to the congestion problem lies in the use of information and automation technology, including the scientific planning and management of the transportation network.

Performance measurement is the first step in intelligent management. This is an area in which information technology can make a big impact. However, existing methods of performance measurement rely largely on manual data collection. Because this is labor-intensive, data are collected infrequently, and do not adequately describe traffic congestion. For example, the annual HICOMP report, the main Caltrans performance measurement

effort, uses probe vehicle data that are collected only twice a year. Annual figures are extrapolated from these few measurement points and are misleading because of the variability in congestion.

Enter PeMS, whose founding principles include automated data collection and processing. PeMS collects data from thousands of detectors in California every 30 seconds. Its database contains terabytes and grows at 2GB per day. The equivalent of the HICOMP report is implemented as an application in PeMS, which can be produced instantaneously on demand. Because it uses a much richer data set than the HICOMP, the PeMS results are much more accurate.

While performance measurement requires a database with large enough coverage, it's not enough to simply publish the raw data. The richness of the information contained in the data can only be realized with proper statistical analysis. PeMS computes many meaningful quantities from the raw measurements of 30-second flow and occupancy. It captures the true range of traffic behavior with statistical descriptions of many phenomena.

There are many ways to implement a data warehousing and data mining system. PeMS is founded on several important principles. We wanted our quantities to be interpretable and reproducible. Specifically, we define and compute these important measures of performance: VMT, VHT, delay, and travel time. The algorithms that compute these quantities are published. The goal is to have a transparent system whose results can be trusted and used for comparison.

PeMS is meant to be widely accessible. This is the reason for its web interface and a suite of menu-driven tools to serve the day-to-day needs of traffic planners and engineers.

It uses as much automation as possible. We collect data from Caltrans TMC's, California Highway Patrol, and the national weather service in real time, and process the data and store the result.

As a policy, PeMS keeps all data on-line, providing a rich historical database for in-depth analysis. Several of these research studies were presented in Chapter 3. These studies on bottleneck capacity, travel time variability, and recurrent/non-recurrent delay demonstrate the wealth of information contained in PeMS, and its potential for understanding and managing congestion. For example, the research results point the way to ramp metering, benefit of ITS, and the effectiveness of Freeway Service Patrol strategies.

PeMS's database presents challenges and opportunities for academic research. Our use of statistical analysis yielded algorithms that compute speed from single loop detectors and predict travel times; we also developed methods to calculate the delay caused by incidents and measure the capacity of bottlenecks. An important part of any data mining system is data integrity. The large number of sensors, maintained by different Caltrans entities, present a difficult challenge of detecting and correcting data errors. From the analysis of historical patterns of data errors and loop failure, we implemented a data quality control layer in PeMS that detects and imputes bad samples. This layer provides trusted data to higher level applications, and insulates analyses from sensor failures. On another front, we also developed on-line tools that help Caltrans maintenance crews to diagnose and fix detectors.

The future of PeMS is bright. Although it has been mostly a research project, PeMS has already proven itself to be a useful tool. We provided performance measures for

reports at the request of Caltrans and the FHWA, as well as news agencies. Our raw and processed data are used by value added resellers who provide traveler information services to the public. Even while PeMS was still a research project, hundreds of Caltrans employees had accounts on PeMS, and used it thousands of times per month. By January 2003, PeMS will be installed at Caltrans as an official, operational system.

At the same time, researchers are using PeMS to provide solutions to practical problems. We are implementing a program to publish predicted travel times and alternate routes on Changeable Message Signs (CMS). We are also working with Caltrans Detector Fitness Program (DFP) to develop a computerized inventory and diagnostics tool for maintenance crews. Many research papers were published using PeMS data and system on topics in speed estimation, performance measurement, traveler information, travel time, and data quality. We believe that PeMS's data and application of statistical analysis and other computer and information technologies will enable scientific management of the transportation system, and promote deeper understanding of traffic phenomena.

Bibliography

[1] "California Highway Patrol incident website," http://cad.chp.ca.gov.

[2] David Schrank and Tim Lomax, "The 2002 urban mobility report," Tech. Rep., Texas Transportation Institute, June 2002, http://mobility.tamu.edu.

[3] Mary K. Teets, "Highway statistics 1996," Tech. Rep. FHWA-PL-98-003, Federal Highway Administration, Washington, DC, 1997.

[4] "Caltrans today," http://www.dot.ca.gov/hq/paffairs/about/today.htm.

[5] M. Theresa Smith, "The impact of highway infrastructure on economic performance," *Public Roads*, vol. 57, no. 3, pp. 8–14, Spring 1994.

[6] "Analysis of the 2002-2003 budget bill," Tech. Rep., Legislative Analyst's Office, 2002, http://www.lao.ca.gov/analysis_2002/transportation/trans_05_deptoftrans\%_2660_anl02.htm.

[7] Mary C. Hill, Ed., *Performance measures for California transportation system users and investors*. Department of City and Regional Planning, UC Berkeley, October 1997.

[8] Pravin Varaiya, "How to measure transportation system performance," http:

//paleale.eecs.berkeley.edu/~varaiya/papers_ps.dir/TSperf.pdf, December 1997.

[9] Highway Operations Caltrans District 4, "1998 highway congestion monitoring report (HICOMP)," Tech. Rep., Caltrans District 4, 1998.

[10] Traffic Operations Program, "1999 traffic volumes on California state highways," Tech. Rep., State of California, Business, Transportation and Housing Agency, Department of Transportation, Sacramento CA, June 2000.

[11] Lawrence A. Klein, "Detection technology for IVHS," Tech. Rep. FHWA-RD-95-100, Federal Highway Administration, McLean, VA, December 1996.

[12] Shawn M. Turner, William L. Eisele, Robert J. Benz, and Douglas J. Holdener, "Travel time collection handbook," Tech. Rep. FHWA-PL-98-035, Federal Highway Administration, Washington, DC, March 1998.

[13] Laurie Blake, "It's September – when traffic again turns ugly," *Minneapolis - St. Paul Star Tribune*, September 22 2002.

[14] E.B. Lee, C.W. Ibbs, J.R. Roesler, and J.T. Harvey, "Construction productivity and constraints for concrete pavement rehabilitation in urban corridors," *Transportation Research Record* 1712, , no. 1712, pp. 13–22, October 2000.

[15] Fred L. Hall and Kwaku Agyemang-Duah, "Freeway capacity drop and the definition of capacity," *Transportation Research Record* 1320, pp. 91–98, 1991.

[16] James H. Banks, "Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering?," *Transportation Research Record* 1320, pp. 83–90, 1991.

[17] C. Ng and C. Chen, "Wireless content delivery and user profiling," in *2001 IEEE Intelligent Transportation Systems Proceedings*, Oakland, CA, 2001, pp. 973–975.

[18] Boris Kerner, "Theory of congested traffic flow: self-organization without bottlenecks," in *Fourteenth International Symposium on Transportation and Traffic Theory*, Jerusalum, Israel, 1999, pp. 147–171.

[19] M. Cassidy and R. Bertini, "Some traffic features at freeway bottlenecks," *Transportation Research B*, vol. 33B, pp. 25–42, 1999.

[20] Paul Neel and Jason Gibbens, "SR520 eastbound morning ramp metering three month study," Tech. Rep., Washington State Department of Transportation, January 2001, http://www.wsdot.wa.gov/Projects/SR520RampMeters/.

[21] Cambridge Systematics Inc, "Twin Cities ramp meter evaluation," Tech. Rep. HF2891, Minnesota Department of Transportation, February 2001.

[22] *Highway Capacity Manual*, TRB, National Research Council, Washington DC, 1994.

[23] C. Chen, Z. Jia, and P. Varaiya, "Causes and cures of highway congestion," *Control Systems Magazine*, vol. 21, no. 4, pp. 26–33, December 2001.

[24] M. Cassidy and A. May, "Proposed analytical technique for estimating capacity and level of service of major freeway weaving sections," *Transportation Research Record* 1320, pp. 99–109, 1991.

[25] Feng-Bor Lin and Cheng-Wei Su, "A methodology for capacity and level-of-service analysis of freeway basic selections," *Transportation Planning Journal*, vol. 27, no. 1, pp. 1–34, May 1998.

[26] K. Small, R. Noland, X. Chu, and D. Lewis, "Valuation of travel-time savings and predictability in congested conditions for highway user-cost estimation," Tech. Rep. 431, National Cooperative Highway Research Program, TRB, National Research Council, Washington DC, 1999.

[27] H. Wakabayashi, "Improvement of terminal reliability and travel time reliability under traffic management," in *Proceedings, Pacific Rim Transportation Technology Conference*, New York, 1993, vol. 1, pp. 211–217, American Society of Civil Engineers.

[28] T. Oda, "An algorithm for prediction of travel time using vehicle sensor data," in *Third International Conference on Road Traffic Control*, London, England, 1990, pp. 40–44, Institution of Electrical Engineers.

[29] Erik van Zwet, "A simple and effective method for predicting travel times on freeways," in *2001 IEEE Intelligent Transportation Systems Proceedings*, Oakland, CA, 2001, pp. 227–232.

[30] A. Skabardonis, K. Petty, H. Noeimi, D. Rydzewski, and P. Varaiya, "The I880 field experiment: analysis of incident data," *Transportation Research Record* 1603, pp. 72–79, 1996.

[31] A. Skabardonis, K. Petty, H. Noeimi, D. Rydzewski, and PP Varaiya, "The I880

field experiment: database development and incident delay estimation procedures," *Transportation Research Record* 1554, pp. 204–212, 1996.

[32] James H. Kell, Iris J. Fullerton, and Milton K. Mills, "Traffic detector handbook," Tech. Rep. FHWA-IP-90-002, U.S. Department of Transportation, Federal Highway Administration, Office of Research and Development, July 1990.

[33] H.J. Payne, E.D. Helfenbein, and H.C. Knobel, "Development and testing of incident detection algorithms," Tech. Rep. FHWA-RD-76-20, Federal Highway Administration, Washington DC, 1976.

[34] L. Jacobson, N. Nihan, and J. Bender, "Detecting erroneous loop detector data in a freeway traffic management system," *Transportation Research Record* 1287, pp. 151–166, 1990.

[35] D. Cleghorn, F. Hall, and D. Garbuio, "Improved data screening techniques for freeway traffic management systems," *Transportation Research Record* 120, pp. 17–23, 1991.

[36] R. E. Turochy and B. L. Smith, "A new procedure for detector data screening in traffic management systems," *Transportation Research Record* 1727, pp. 127–131, 1991.

[37] G. Davis and N. Nihan, "Using time-series designs to estimate changes in freeway level of service, despite missing data," *Transportation Research. Part A*, vol. 18A, no. 5/6, pp. 431–438, October 1984.

[38] D. Dailey, "Improved error detection for inductive loop sensors," Tech. Rep. WA-RD 3001, Washington State DOT, May 1993.

[39] D. Dailey, "A statistical algorithm for estimating speed from single loop volume and occupancy measurements," *Transportation Research B*, vol. 33B, no. 5, pp. 313–322, June 1999.

[40] Z. Jia, B. Coifman, C. Chen, and P. Varaiya, "The PeMS algorithm for accurate, real-time estimates of g-factors and speeds from single loop detectors," in *2001 IEEE Intelligent Transportation Systems Proceedings*, Oakland, CA, 2001, pp. 527–532.

[41] W.S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[42] Jin Y. Yen, "Finding the K shortest loopless paths in a neetwork," *Management Science*, vol. 17, no. 11, pp. 712–716, July 1971.

[43] E. Lawler, "A procedure for computing the K best solutions to discrete optimization problems and its application to the shortest path problem," *Management Science*, vol. 18, no. 7, pp. 401–405, March 1972.

Index