Creating transportation system intelligence

Pravin Varaiya
University of California, Berkeley

Information technology (IT) provides the means to store, manipulate and disseminate massive amounts of data. The integration of IT at all levels of the transportation system creates the “intelligence” in Intelligent Transportation Systems (ITS). But this integration is a long and difficult process of the search for, and the exploitation of, the numerous opportunities in the interconnected set of operations, planning and investment procedures constituting today’s transportation systems.

This paper gives a glimpse into the opportunities for enhancing freeway systems productivity. The discussion summarizes three years of experience with the Performance Measurement System or PeMS—a system that collects and stores data from California loop detectors, and converts these data into useful information. Examples from Los Angeles illustrate how this information can improve system management, assist travelers, and challenge current understanding of freeway traffic behavior.

With PeMS, freeway operations and planning can be based on information that previously was unavailable or too costly to gather. Routine reports, like California’s congestion monitoring report that today consume appreciable resources to produce, can be generated at no cost. Significantly, engineers and planners can quickly isolate problem areas and focus on potential solutions as seen in two examples: identification of bottlenecks, and location of freeway segments where intelligent ramp-metering can significantly reduce congestion.

Travelers face a large variation in travel time during peak hours. Less appreciated is the fact that knowledge of the current state of the traffic can drastically reduce this variation, so travel times can be accurately predicted. PeMS makes these predictions available, and uses them to suggest optimum routes.

Our understanding of traffic behavior is embodied in models that inform decisions and professional training. Typically in traffic engineering, these models are insufficiently validated. However, PeMS data can be easily used to test these models, as seen in two surprising findings. First, maximum throughput occurs at the free flow speed of 60 mph, and not between 35 and 50 mph. Second, a large fraction of freeway congestion delay is due to inefficient operation rather than to excess demand.

ITS is often simply associated with a set of technologies (AVL, CMS, ETC, VIPS, etc), to be deployed one at a time in return for corresponding incremental benefits. But the paper suggests that much larger productivity gains in the freeway system can be obtained when operations, planning, and investment processes are transformed to make intelligent use of these technologies guided by the kinds of information that PeMS provides.

The paper begins with a brief overview of PeMS and then gives examples of PeMS applications. The reader will gain a better appreciation of PeMS from its website http://transact.eecs.Berkeley.EDU.

System overview

PeMS collects and stores data from loop detectors operated by the California Department of Transportation (Caltrans). PeMS applications convert these data into useful information accessed through the Internet by Caltrans personnel, value-added resellers (VARs), the public, and the research community. PeMS is a functioning prototype. It will be deployed statewide in July 2002. It is a low-cost system, built from commercial, off-the-shelf components. PeMS can be deployed incrementally, with no disruption of existing procedures. (See reference 1.) Its software is open: other data sources will be incorporated in PeMS as they become available; furthermore, PeMS is designed to include electronic data from other transportation modes, such as transit.

The PeMS database computer is located in the University of California, Berkeley. The computer has 4 GB of main memory, and 4 terabytes of disk—enough to store several years of California data online.

The software is organized in three layers. At the bottom is database administration. The work is standard but highly specialized: disk management, crash recovery, table configuration. The middle layer comprises software that works on the data as they arrive in real time. It

- Aggregates 30-second flow and occupancy values into lane-by-lane, 5-minute values;
- Calculates the *g*-factor for each loop, and then the speed for each lane. (Most detectors in California are single loop, and only report flow and occupancy. PeMS adaptively estimates the *g*-factor for each loop and time interval. The algorithm and tests of its validity are reported in reference 2.);
- Aggregates lane-by-lane values of flow, occupancy, and speed across all lanes at each detector station. At this point, PeMS has flow, occupancy, and speed for each 5-minute interval for each detector station (one station typically serves the detectors in all the lanes at one location);
- Computes basic performance measures such as congestion delay, vehicle-miles-traveled, vehicle-hours-traveled, and travel times.

The top software layer comprises applications some of which are described below.

Routine reports

Systems at Caltrans depend on monthly or annual reports and programs that provide high-level information to policy makers. These include the Traffic Operations Strategies report, the Highway Congestion Monitoring Program (HICOMP), and System Performance Measures Initiative. PeMS can benefit these reports and programs, as seen in the case of the annual HICOMP report.

The HICOMP report presents the location, magnitude, and duration of congestion for California freeways. Caltrans uses this information to identify problem areas and to establish priorities for operational and air quality improvement projects. Data for the report are obtained from “floating” cars driven through 5-7 mile freeway sections twice a year during congested periods. Figure 1 is one of PeMS’ “average plots.” It gives the maximum, average, and minimum vehicle-hours of congestion delay—the extra time spent driving below 35 mph—on US-101N for each day of the week, averaged over a 16-week period beginning February 4, 2001.

During these 16 Wednesdays (day 4 in Figure 1), the delay ranged from a minimum of 10,000 to a maximum of 60,000 vehicle-hours. This 600 percent variation implies that the twice-a-year samples of the HICOMP report are unreliable. With PeMS it is possible to track congestion accurately to determine trends and departures from the trend to produce (at no cost) a report that gives meaningful statistical measures of congestion. More significantly, PeMS forces the recognition that congestion delay is a random quantity so its measurement and report must take the statistical fluctuations into account.

The preceding remark applies equally to many reports (such as census counts) based on one-shot samples of randomly fluctuating quantities. Travel time is an important component “mobility” measures. PeMS computes the travel time for each freeway segment starting every five minutes. As shown later, travel times fluctuate widely from day to day. A meaningful summary of travel times must reflect these fluctuations.

PeMS also collects statewide incident data reported by the California Highway Patrol, and locates these incidents on the same geographical basis as the loop detector data. So hypotheses relating incidents to traffic variables such as vehicle-miles traveled or congestion delay or freeway geometry can be formulated and tested. Today, the relation between incidents and congestion delay is based on folklore. PeMS permits the discovery and accurate specification of such a relation, if one exists.

Finding bottlenecks

The PeMS application, called “plots across space,” can assist in identifying bottleneck locations for more detailed investigation. To use the application, the engineer selects a section of freeway, a time, and a performance variable such as speed, flow, or delay. PeMS returns a plot of the variable across space. The plot in Figure 2, for example, displays speed averaged across all lanes in the figure—PeMS also provides lane-specific plots—over a 30-mile stretch of I10-W, beginning at post mile 20, at 7.30 am on September 14, 2000.

The precipitous drop in speed from 60 to 20 mph near post mile 23 indicates a potential bottleneck. There is another potential bottleneck near post mile 32. The PeMS contour plot of Figure 3 confirms the existence of both bottlenecks. Further confirmation may be obtained by examining the same plots for other days. Without PeMS this would be a very time-consuming analysis. Having quickly determined the existence of these bottlenecks, the engineer can go on to determine their cause, such as the location of interchanges, the highway geometry, large flows at ramps, etc, and propose potential solutions to alleviate the bottleneck.

Observe how PeMS can dramatically shift the use of the engineer's time from the drudgery of collecting data and making contour plots, to the creative understanding of the causes of bottlenecks and finding opportunities for relieving them. Furthermore, any scheme implemented to relieve a bottleneck can be rigorously evaluated by a thorough before-and-after comparison. Different schemes can be compared. Over time, there will be an accumulation of experience statewide so that schemes can be implemented with some degree of confidence in their effectiveness.

Maximum flow occurs at 60 mph

The speed-flow relationship is fundamental to traffic theory. In the Highway Capacity Manual this relation is given by a smooth curve, which yields a maximum flow at a speed between 35 and 50 mph. We use PeMS to test this hypothesis using cross-sectional data from all 3,363 functioning loop detectors (out of a total of 4,199 detectors) at 1,324 locations in Los Angeles, for a 12-hour period beginning midnight of September 1, 2000, bracketing the morning commute hours.

For each detector we find the 5-minute interval during which the flow reaches its maximum value. We then calculate the average speed during a 25-minute interval surrounding this maximum-flow interval. This is a measure of the *sustained* speed at the time of maximum flow. Figure 4 displays the per-lane distribution of this speed: in (the innermost) lane 1 this speed is between 60 and 70 mph, in lane 2 it is between 55 and 60, in lanes 3 and 4 between 50 and 60. The test rejects the hypothesis that maximum flow occurs between 35 and 50 mph.

The finding has some important implications. First, congestion delay should be measured as the time spent driving below 60 mph, both because it is the most efficient speed and because drivers experience congestion below this speed. (Caltrans today measures congestion as the time spent driving under 35 mph continuously for 15 minutes.)

Second, a ramp-metering strategy will only be effective if it maintains free flow speed. Lower speeds, such as say 45 mph, are simply not sustainable. Figure 5 illustrates this. (It is an instance of a PeMS "x-y plot" application in which 5-minute averages of any two variables at a detector are plotted against one another.) It gives the speed-flow relationship on lane 1 between 4.00 and 8.00 am on September 14, 2000, at post mile 32.87 on I-10W, near the second bottleneck in Figure 2. Clearly, once occupancy increases to cause the speed to drop below 60 mph at 5.10 am, the flow becomes unstable, dropping to 30 mph by 5.30, and 15 mph by 7.00 am. It seems unlikely that traffic flow at any speed below 55 mph can be sustained. This conclusion is confirmed by examining hundreds of similar plots.

Potential gains from ramp-metering

A fairly complex PeMS application calculates the potential reduction in congestion from an ideal ramp-metering policy (IMP). We study a freeway section that experiences recurrent congestion during the morning rush hour from 6.00 to 10.00 am, on a particular day. In the example this is the 16-mile segment of I-210W, starting at post mile 22, on January 11, 2001. The time period spans the rush hour, say 4.00 am to noon. So traffic is free flowing at the beginning and end of the study period.

The freeway section comprises several PeMS segments. (A segment is the freeway surrounding a detector station half way to the two neighboring stations). Some segments have on-ramps, some have off-ramps, and some have neither. PeMS gives, for each 30-second interval, the inflows of vehicles into the study section from each on-ramp and from upstream of the section, as well as the outflows at each off-ramp and downstream of the section. PeMS does not have origin-destination data, so the application calculates the constant turning ratio that matches total inflow and outflow in each segment.

The application next calculates the maximum throughput in each segment. This is simply the maximum flow that was in fact observed in that segment during the study time period, 4.00 am to noon, January 11, 2001. The maximum flow is an empirical quantity, which varies slightly from day to day and from segment to segment, see reference 4.

The hypothesis underlying the application is this: if the flow on each segment is *always* maintained below this maximum (say, by 3 percent), then vehicles on that segment will travel at 60 mph. The preceding section strongly supports this hypothesis, although a true test would require field experiments.

IMP imposes the policy that at each on-ramp (and upstream of the study section) vehicles are admitted so long as the flow in every section does not exceed the maximum flow, less 3 percent. (This is not the way to implement IMP; that should be based on measuring occupancy downstream of each on-ramp.) With this policy, under the hypothesis above, a vehicle may be held back at an on-ramp, but once it enters the freeway it will travel at 60 mph.

The result of IMP is displayed in the three plots of Figure 6. The top curve plots the *actual* vehicle-hours spent in the study section during each 5-minute slice from 4.00 am to noon. (The units are normalized to vehicle-hours/hour. This is a simple calculation since flows and speeds are known.) So the area under the top curve is the total vehicle-hours actually spent in the section during that period. The bottom curve gives the vehicle-hours those vehicles would have been spent if they experienced no delay at the ramps and traveled at 60 mph. So the area under the bottom curve is the *free flow* vehicle-hours that would have been spent by the same traffic demand. The difference in the area under the top and bottom curves is the vehicle-hours of delay suffered by traveling less than 60 mph. The two curves coincide outside the 6.00 to 10.00 am congestion period, as expected.

The middle curve plots the vehicle-hours that would have been spent under the IMP metering policy. Recall that under the hypothesis above, a vehicle is either queued up at an on-ramp or traveling at 60 mph on the freeway. So the area between the middle and bottom curves is the vehicle-hours spent queued up at the on-ramps, and the area between the top and middle curves is the net reduction in congestion delay. In the example of Figure 4, there is about 3,000 vehicle-hours of total congestion delay, of which 2,400 is eliminated by IMP with 600 vehicle-hours of queuing delay at on-ramps.

This application can be used by planners to locate potential sites where ramp-metering may be advantageous. Ramp-metering is a contentious local public policy issue in California, and most discussion is based on unfounded allegations about its impact. PeMS can provide an empirical basis for estimating the cost and benefit of a proposed ramp-metering scheme. The application also calculates the queue lengths formed at all the ramps and upstream of the study section. That information can be used to determine whether there is sufficient capacity in the ramps. As is known, the queue at one on-ramp can be traded off against another. So the application can stimulate a study of alternative coordinated ramp-metering strategies, and coordinated arterial signaling.

We use this application for five freeways in Los Angeles for the morning commute periods of the week of October 3-9, 2000. We then “blow up” the results to all LA freeways. This leads to the estimate that travelers spend an extra 70 million vehicle-hours each year driving below 60 mph, of which 50 million-hours can be eliminated by intelligent ramp-metering. At (say) $20 per vehicle-hour, there is a potential annual savings of $1 billion, see reference 3.

Efficiency of freeway operations

The freeway segment of Figure 5 can support a flow of 2,100 vehicles/lane/hour at 60 mph. But at 7.00 am, when congestion is worst, it serves only 1,300 vehicles/lane/hour at 15 mph, indicating a drop in operating efficiency. We propose a measure of efficiency, η, given by the formula

$$\eta = \frac{Flow \times Speed}{MaxFlow \times SpeedAtMaxFlow(60)} \ . \tag{1}$$

According to this formula, the efficiency of this segment at the time of worst congestion was $\eta = \frac{1300 \times 15}{2100 \times 60} = 13$ percent.

We justify formula (1) by viewing the freeway segment as a queuing system. The queuing system provides a service to each customer (vehicle)—the transport of the vehicle across the segment. The vehicle’s service time is $\frac{SegmentLength}{Speed}$. The system serves $Flow$ vehicles in parallel. The *throughput* of this queuing system at any time is the number of vehicles served per hour at that time, namely $\frac{Speed}{SegmentLength} \times Flow$. The maximum throughput is $\frac{SpeedAtMaxFlow(60)}{SegmentLength} \times MaxFlow$. Formula (1) defines efficiency as the ratio of actual throughput to maximum throughput.

We use PeMS to estimate the efficiency of all 291 segments of I-10W with functioning detectors during the morning congestion period on October 1, 2000. For each segment we find the 5-minute interval between midnight and noon when its detector recorded the maximum occupancy. This is the time of worst congestion, and we find the speed and flow at this time. As before, we also find the maximum flow during the 12-hour interval. Using these in formula (1) gives the efficiency of the segment during worst congestion.

Figure 7 gives the distribution of efficiency for the 291 segments at the time of worst congestion: 78 segments had efficiency under 40 percent, 65 had efficiency between 40 and 80 percent, 71 had efficiency between 80 and 100 percent, and 46 had efficiency above 100 percent (they recorded a speed above 60 mph at time of maximum occupancy).

California's freeway system cost $1 trillion dollars. The calculation above shows that this capital stock is operated at a very low efficiency precisely at times of greatest demand (worst congestion). The potential gain from restoring efficiency will far exceed any practically conceivable increases in capacity through new construction. Any program to build "intelligence" in the transportation system must surely take as its main objective the recovery of this efficiency loss.

Travel times

Figure 8 gives the travel times for a 48-mile trip over I-10E, beginning at post mile 1.3 at any time between 5 am and 8 pm, for 20 working days in October 2000. The data are obtained from PeMS' travel time calculations.

Anyone planning this trip faces the statistical distribution implicit in this figure. If you intend to leave at say 5 pm, the "vertical slice" through the figure at 5 pm gives the prior distribution you face. So your trip may take between 45 and 130 minutes, with a 70 percent chance it will take between 60 and 100 minutes and a 10 percent chance it will take more than100 minutes. If you want to place a 90 percent confidence interval around your travel time, the best you can do is between 55 and 110 minutes—a 200 percent variation. But this variation can be drastically reduced if you know current conditions.

There are 20 curves in the figure, one for each day. The curve for a particular day is obtained from the travel times for that day starting every 5 minutes between 5 am and 8 pm. It is evident that if the trip starting at 5 pm takes more than 100 minutes, then that trip belongs to curves (days or random draws) for which the trip starting at 4 pm takes more than 90 minutes. What this means is that if you know the *current* travel time for a particular trip, you can predict the *future* travel time quite well. If at 4 pm the travel time is 90 minutes, then you can be 90 percent confident that at 5 pm the travel time will be between 85 and 110 minutes—a 25 percent variation. On the other hand, if at 4 pm the travel time is 60 minutes, the 90 percent confidence interval for a trip staring at 5 pm is between 60 and 80 minutes. In statistical terms this means that the unconditional variance in the travel time distribution is large, but the variance on the distribution *conditioned* on present and past values is much smaller. Of course, the further into the future you want to predict, the less important is knowledge of the current state of traffic: if you know the travel time at 5 am, and you want to predict the travel time at 5 pm, you can't do better than the unconditional distribution.

A PeMS application makes point estimates of future travel times for each freeway segment based on current and past travel times. (The past travel times are stratified by day of week and time of day to extract trends. See reference 5.) Through her browser, a user first indicates a proposed trip by clicking on the origin and the destination on the freeway map. The user then selects a start time (or the arrival time) and PeMS calculates 15 routes and their travel time estimates, including the routes with the shortest travel time and the shortest distance. The other routes differ by one link from these.

Other applications

The preceding sections indicate some uses of PeMS. We quickly list some others. Freeway lanes are often closed in response to requests for scheduled maintenance. With PeMS, one can compute the likely delay caused by a proposed lane closure, by comparing traffic demand for similar time intervals in the past with the reduction in throughput from the lane closure. So proposed lane closures may be shifted to time intervals to minimize the impact. A more ambitious scheme might involve including incentives in maintenance contracts that reflect these delays.

PeMS collects data on HoV lanes. These data can be used to determine the shift to car-pooling as a function of the congestion in the mainline lanes. If ramp-metering eliminates mainline congestion as suggested above, HoV lanes offer no advantage, and can be converted to mainline lanes, thereby increasing capacity. HoV bypass lanes at on-ramps can nonetheless encourage car-pooling.

General-purpose simulation models like Corsim and Paramics are typically used to answer a large number of “if-then” questions ranging from the effectiveness of ramp-metering schemes to the impact of traveler information. These models have many parameters and PeMS data may be used to calibrate those models. A more useful direction of research is the use of PeMS data to create a number of special-purpose statistical models. Statistical models of the impact of lane closures and HoV effectiveness are examples, as are the models for travel time prediction. Other models might estimate impact of incidents, weather, and special events. These special-purpose statistical models are easier to calibrate and maintain, and, where applicable, more reliable than the general-purpose simulation models.

Concluding remarks

Over the past 20 years, there has been a dramatic increase in productivity in the manufacturing sector. Much of this increase can be attributed to the integration of Information Technology (IT) into manufacturing. This has been a long process of learning, trial and error, seeking out opportunities for improvement. The process has been painful as Darwinian selection weeded out firms that did not successfully exploit the opportunities opened up by IT.

ITS initially was greeted as a set of technologies (AVL, CMS, ETC, VIPS, etc) that promised a quick path to productivity gains in transportation. A realistic assessment of actual productivity gains from the deployment of these technologies has not been made (there are several *un*realistic assessments), but an informed guess is that those gains are marginal. In retrospect this is hardly surprising. Like manufacturing, the production of transportation services is a highly complex operation, involving the orchestration of numerous interdependent activities, conventionally classified as operations, planning, and investment. Such highly complex systems do not admit quick technological fixes.

A wise traffic engineer remarked 40 years ago, "If you don't know how your system performed yesterday, you cannot expect to manage it today." A prerequisite to intelligent transportation systems is intelligence—the knowledge of what is happening to the system, an understanding of what decisions are effective, what key opportunities there are in the "value chain" that produces transportation services, and what technologies can help exploit them. Most transportation agencies operate without this intelligence.

It is instructive to think of the freeway system as an agency that consumes resources in order to produce a useful service. The service is transportation—the movement of vehicles carrying people and goods from one place to another. The service produced can be easily measured in vehicle-miles traveled or VMT. The resources consumed by the system are also easily measured. There are the fixed costs: depreciation of the freeway system's capital stock and the (essentially) fixed cost of the workforce that runs the system. There are also the variable costs: the time and money spent by people as they drive the vehicles over the freeways to go from one place to another. This is easily measured, too, as the vehicle-hours traveled or VHT. The ratio, $Q = VMT/VHT$, measures the freeway system's (marginal) productivity.

It is the common experience of drivers in California's urban areas that this productivity is declining. The decline would be much worse were it not for urban sprawl—people and businesses leave areas with very low productivity (this is an explanation of, not an argument for, sprawl); and because people adapt—they change their time of work, use their cell phones, listen to CDs, and give in to "road rage." Agencies that operate freeways usually form a monopoly (excepting a few private toll roads and alternative transit options), their frustrated customers have nowhere else to go, and so productivity continues to decline.

But the status quo can be changed. The first step is to equip system operators and their customers with intelligence about their system. PeMS shows that this step is easy. The next steps are difficult. They require carefully examining all the activities that affect the productivity measure Q, finding the changes in the activities that can lead to the greatest increase in Q, and implementing and monitoring those changes. Making routine the discovery and exploitation of opportunities needs organizational changes within the transportation agency. In the absence of any Darwinian mechanism that rewards those who carry out the required changes, these changes require inspired leadership.

Acknowledgements

This paper reports work of the PeMS Development Group, particularly Chao Chen, Zhanfeng Jia, Karl Petty, Jun Shu, Alex Skabardonis, and Erik van Zwet. I am also grateful to Hamed Benouar, Tom Choe, Ali Grabell, Joe Hecker, Randell Iwasaki, and John Wolf of Caltrans for their interest, support, and criticism of the work of PeMS. Elizabeth Stoltzfus and Chris Williges of Booz.Allen and Hamilton have consistently provided thoughtful comment. Patrick Conroy and Steve Shladover of California PATH program have helped with their encouragement. The research summarized here is supported by grants from Caltrans and the National Science Foundation. The opinions expressed here and the errors of interpretation are those of the author alone.

References

1. C. Chan, K. Petty, A. Skabardonis, P. Varaiya, Z. Jia, “Freeway Performance Measurement System: Mining Loop Detector Data,” 80th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan 2001.

2. Z. Jia, C. Chen, B. Coifman, P. Varaiya, “The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors,” IEEE 4th International ITS Conference, Oakland CA, August 2001.

3. Z. Jia, P.Varaiya, C. Chao, K. Petty, A. Skabardonis, “Maximum throughput in LA freeways occurs at 60 mph v.4,” http://transacct.EECS.Berkeley.EDU/info.phtml, January 2001.

4. Z. Jia, P. Varaiya, C. Chen, K. Petty, A. Skabardonis, “Congestion, excess demand, and effective capacity in California freeways,” http://transacct.EECS.Berkeley.EDU/info.phtml, December 2000.

5. J. Rice and E. van Zwet, “A simple and effective method for predicting travel times on freeways,” IEEE 4th International ITS Conference, Oakland CA, August 2001.

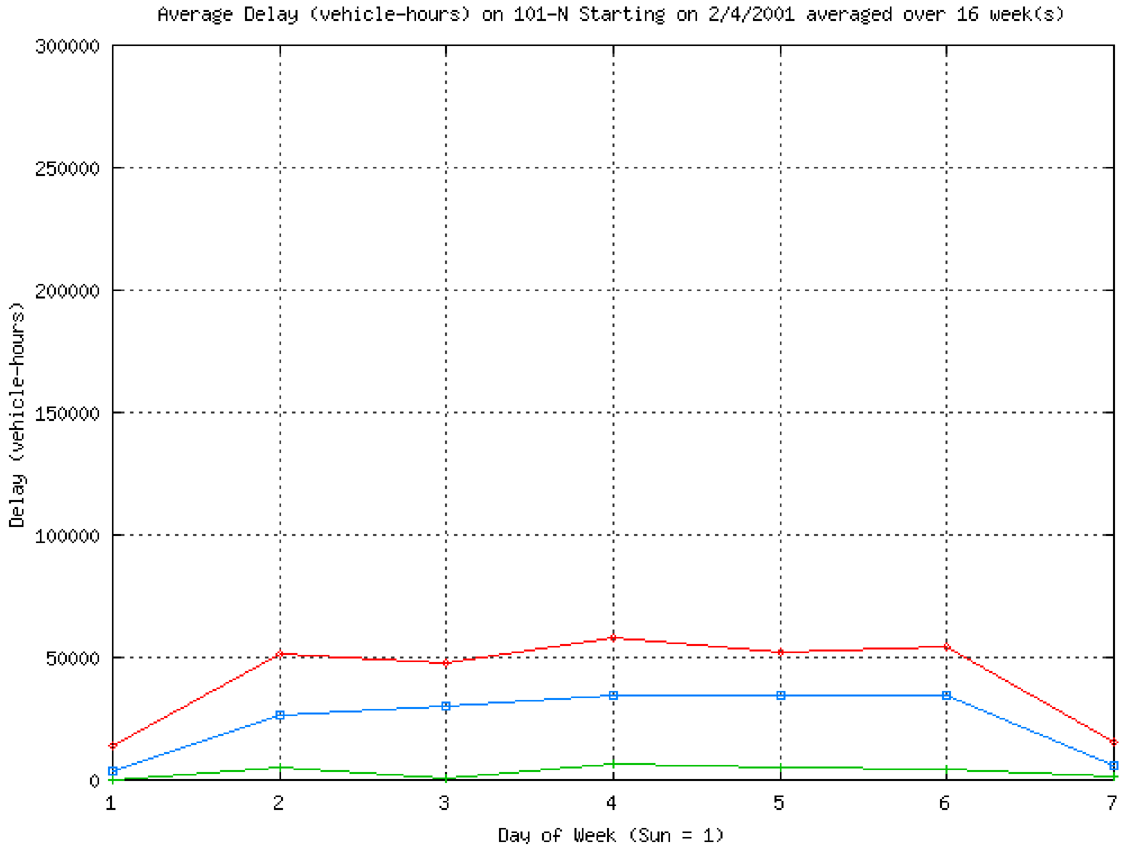

Figure 1 Maximum, minimum, and average delay over US 101-N by day of week

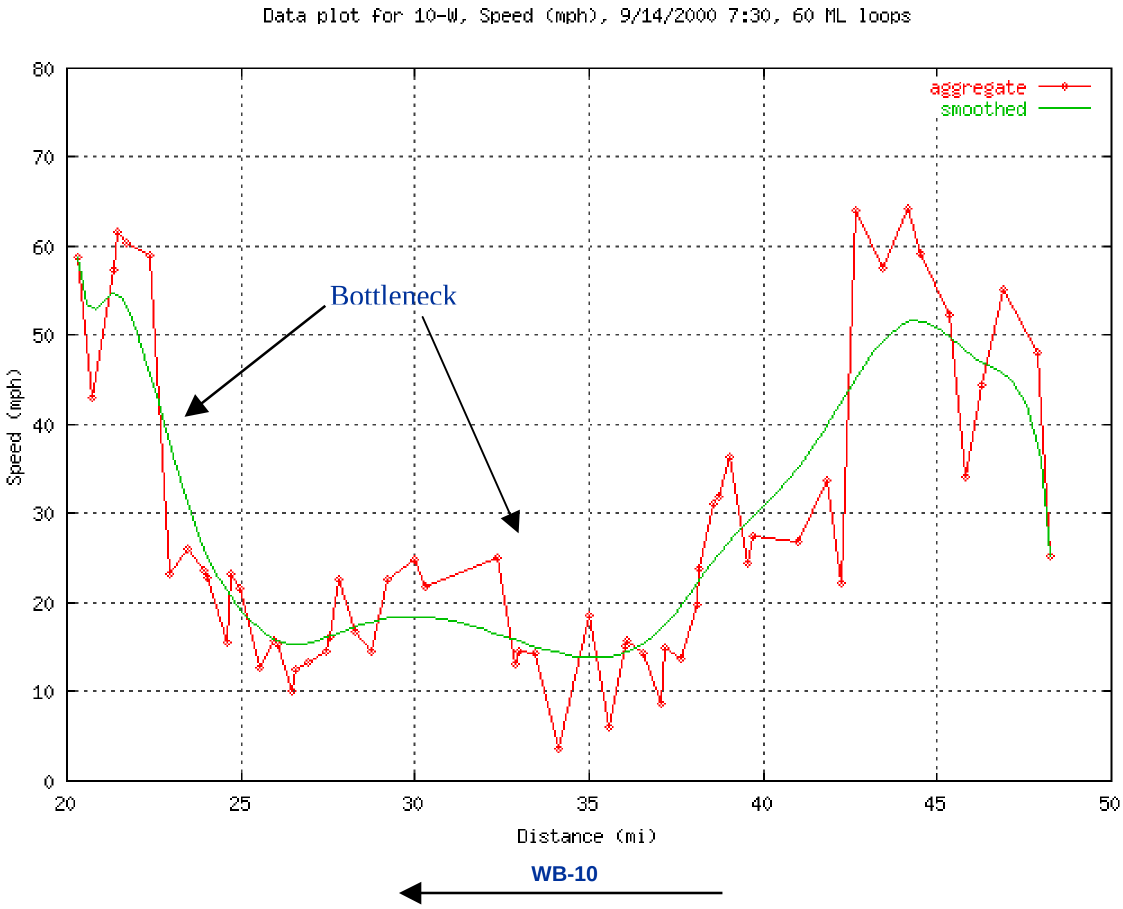

Figure 2 Speed across a 30-mile stretch of I-10W on September 14, 2000 at 7.30 am

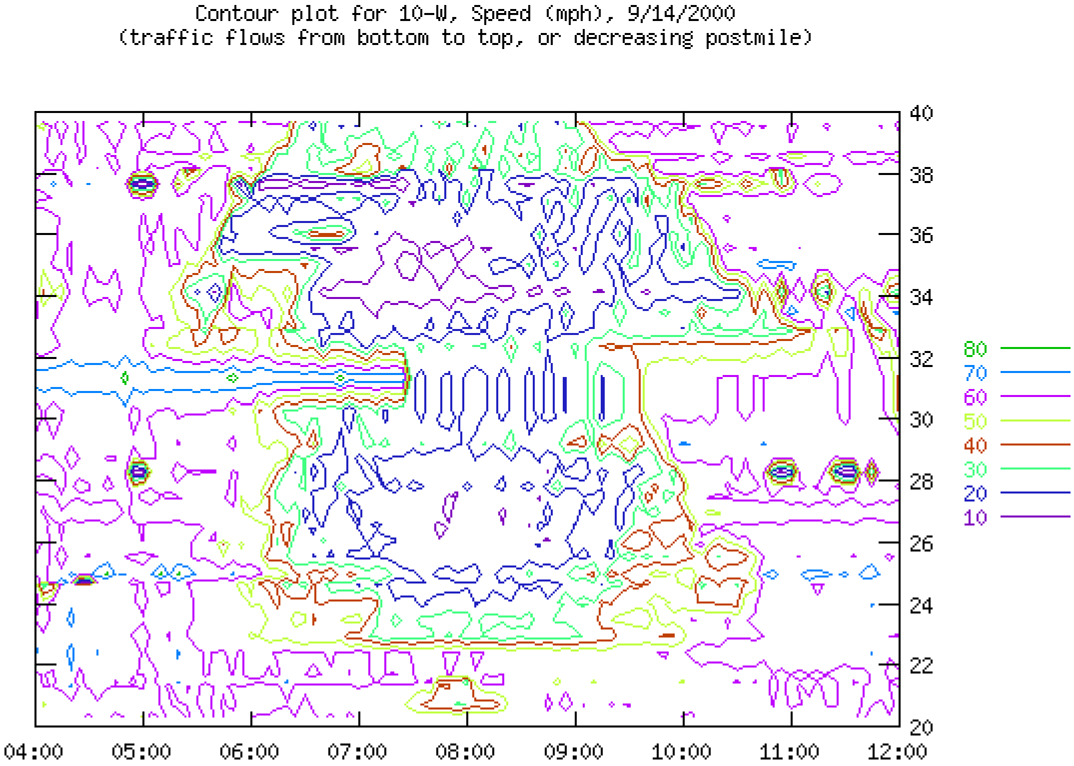

Figure 3 Contour plot of speed on I-10W from 4.00 am to noon, September 14, 2000

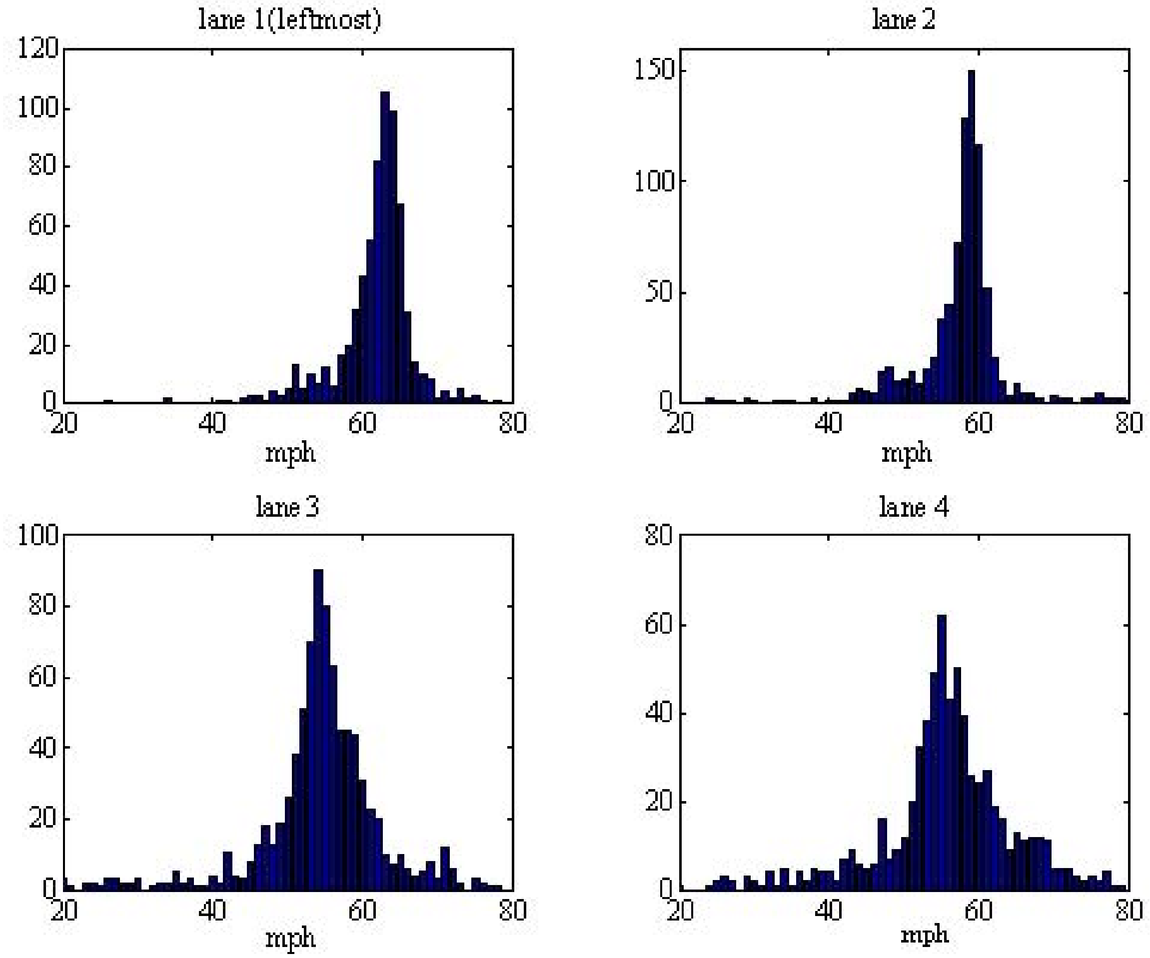

Figure 4 Distribution by lane of detector speed at time of maximum flow

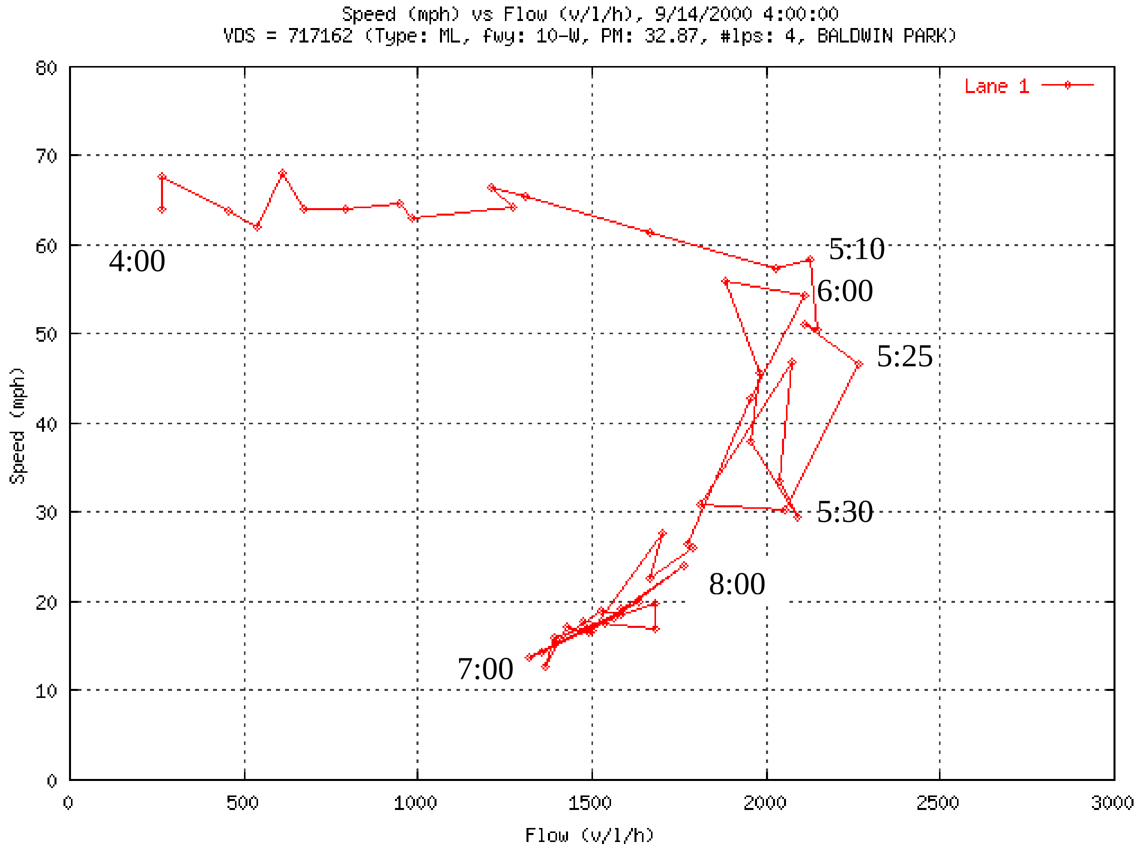

Figure 5 Speed-flow relationship between 4.00 and 7.00 am

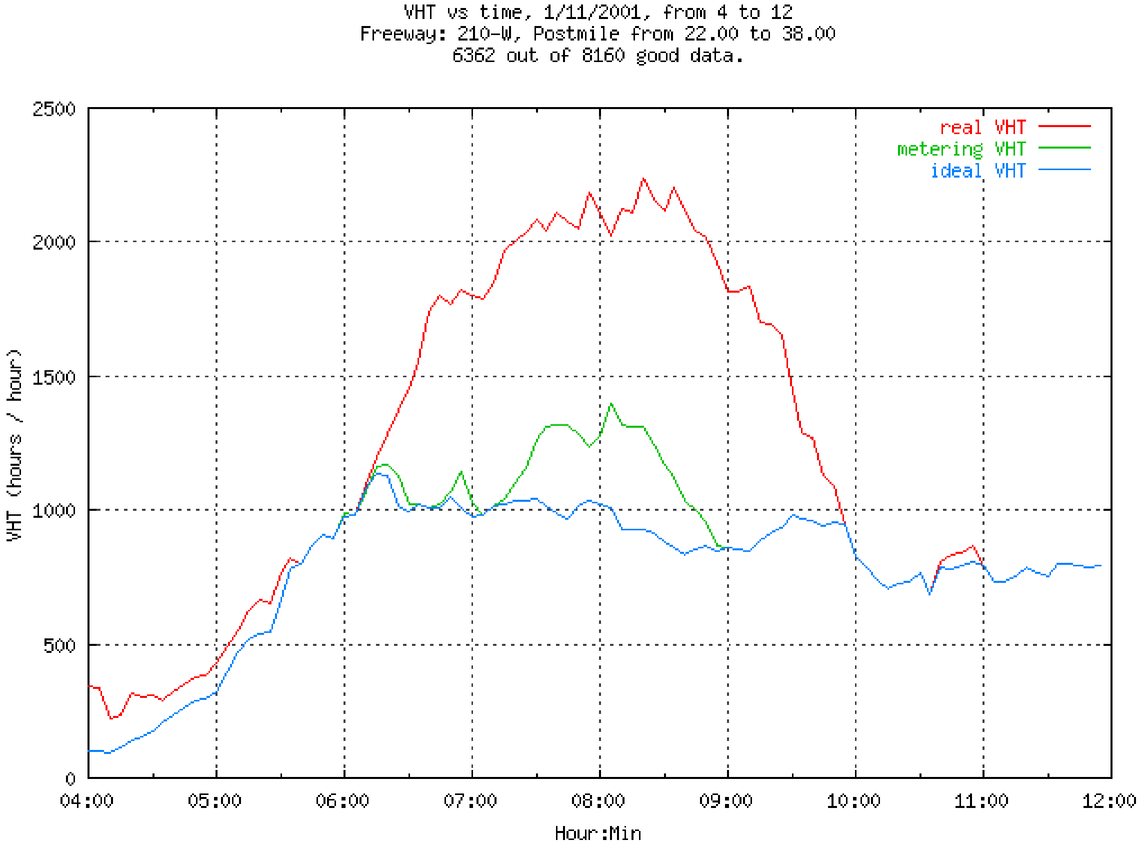

Figure 6 Calculation of potential reduction in congestion delay from ramp-metering

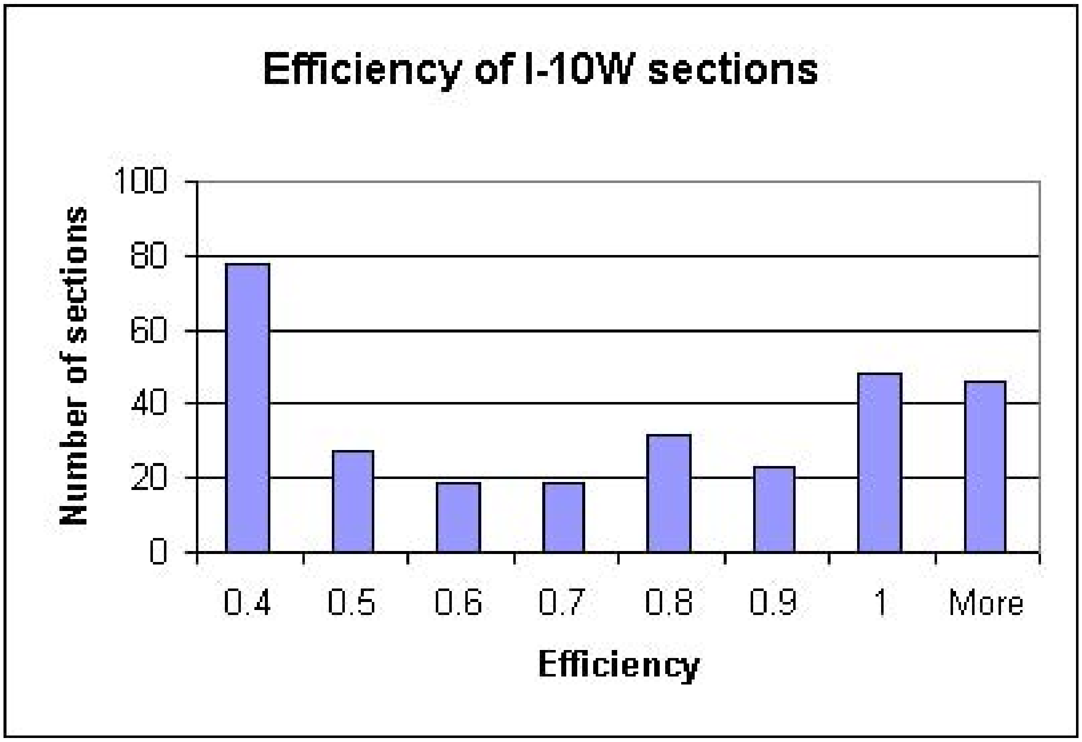

Figure 7 Variation in efficiency during worst congestion along segments of I-10W, midnight-noon, October 1, 2000

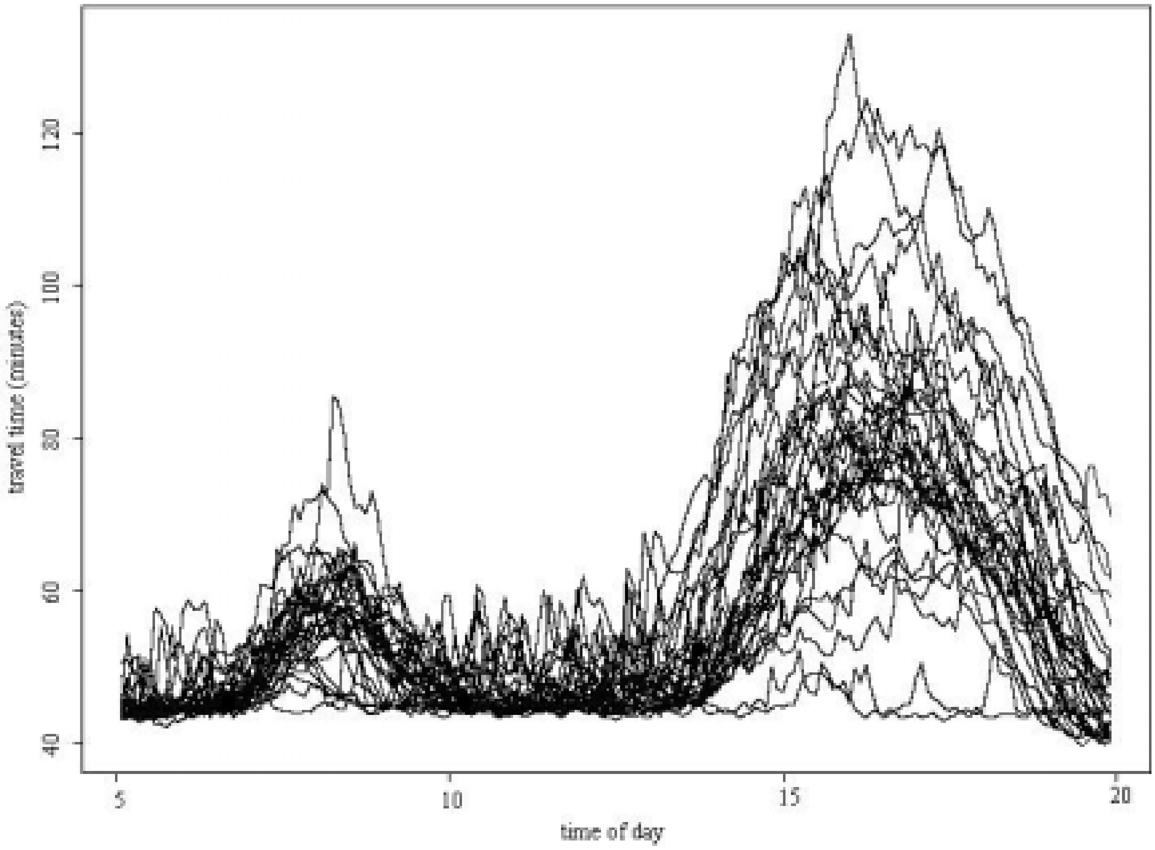

Figure 8 Travel time for I-10E between post miles 1.3 and 48.5 for 20 working days in October, 2000 for different starting times between 5 am and 8 pm