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

Optimal Sensor Requirements

Xuegang (Jeff) Ban, et al.

California PATH Research Report
UCB-ITS-PRR-2009-36

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.

Final Report for Task Order 6328

August 2009

ISSN 1055-1425

california center for innovative transportation
UNIVERSITY OF CALIFORNIA, BERKELEY

FINAL REPORT

PATH TASK ORDER 6328

OPTIMAL SENSOR REQUIREMENTS

Prepared for:

CALTRANS DIVISION OF RESEARCH AND INNOVATION

CALTRANS DIVISION OF TRAFFIC OPERATIONS

Xuegang (Jeff) Ban, Rensselaer Polytechnic Institute, banx@rpi.edu
Alexandre Bayen, University of California, Berkeley, bayen@berkeley.edu
Lianyu Chu, California Center for Innovative Transportation, lchu@calccit.org
Adam Danczyk, University of Minnesota, danc0010@umn.edu
Juan-Carlos Herrera, University of California, Berkeley, jcherrera@berkeley.edu
Ryan Herring, University of California, Berkeley, ryanherring@berkeley.edu
Henry X. Liu, University of Minnesota, henryliu@umn.edu
J.D. Margulici, California Center for Innovative Transportation, jd@calccit.org
Olli-Pekka Tossavainen, University of California, Berkeley, optossav@berkeley.edu
Daniel Work, University of California, Berkeley, dbwork@berkeley.edu

CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION

UNIVERSITY OF CALIFORNIA BERKELEY · 2105 BANCROFT WAY, SUITE 300 · BERKELEY, CA 94720-3830

PHONE: (510) 642-4522 · FAX: (510) 642-0910 · HTTP://WWW.CALCCIT.ORG

PROJECT SUMMARY

Title:	Optimal Sensor Requirements
Sponsor:	Caltrans Division of Research and Innovation
Project Organization:	California Center for Innovative Transportation 2105 Bancroft Way Berkeley, CA 94720 Phone: (510) 642-4522 – Fax: (510) 642-0910
Principal Investigator:	Alexandre M. Bayen, University of California, Berkeley (510) 642-2468 - bayen@ce.berkeley.edu
Project Manager :	JD Margulici, Associate Director, CCIT (510) 642-5929 – jd@calccit.org
Additional Investigators:	Xuegang (Jeff) Ban, Rensselaer Polytechnic Institute Lianyu Chu, California Center for Innovative Transportation Henry X. Liu, University of Minnesota Olli-Pekka Tossavainen, University of California, Berkeley
Center Director:	Thomas West
Administrative Officer:	Coralie Claudel (510) 642-5579 – coralieclaudel@calccit.org
Performance Period:	January 2007 – December 2008
Project Cost:	$299,999

EXECUTIVE SUMMARY

PATH Task Order 6328 addresses the optimal deployment of traffic detectors on freeway to ensure that adequate information is collected at the lowest possible cost. The project team produced a study framework and tools that can be applied locally to test the sensitivity of traffic data quality to detectors location and spacing, and ultimately recommend a deployment plan.

Various types of traffic detectors, including loop detectors, radars, toll tag readers and video cameras are deployed on highways. They provide the data needed to run traffic management applications such as ramp metering control, bottleneck identification, and travel times estimation. However, few studies have systematically analyzed the data requirements of these applications in terms of detector spacing and location. In other words, the trade-offs between the cost of detectors and their benefits for traffic estimation accuracy are not well known. As a result, most highway detectors are installed using ad hoc guidelines or on a case-by-case basis, rather than through the application of measurable objectives. This in turn makes it difficult for practitioners to justify equipment and maintenance expenditures, often slowing deployment.

The product of this research is two-fold. First, we developed a framework to study the sensitivity of traffic information to sensor location and spacing and reached general conclusions. Second, the team created practical tools to assist practitioners at the local level with optimal sensor deployment. These tools include recommendations for rural areas and an Excel-based model for urban areas.

The study framework comprises:

- Quality measures for selected traffic applications, so that information quality can be objectively quantified and various sensor configurations can be benchmarked accordingly;
- An extensive range of study corridors for which highly detailed flow data is available, either through simulation or from cutting-edge traffic experiments. Hypothetical corridors and scenarios can be automatically generated into Paramics microscopic simulation models;
- A sensor model to simulate data capture for a given traffic flow, including a model of detection inaccuracies;
- Tools to process data generated by the sensor model according to various traffic estimation algorithms, most notably a naïve method and a method based on the Cell Transmission Model (CTM);
- Optimization techniques and tools to rapidly identify the best sensor configuration for a given traffic flow, objective function and data processing method;

The results of the investigations conducted within this framework were organized into two sets. Core strategies present first-level analysis, while advanced strategies add refinements to take into account a broader set of objectives and constraints. The results lead to the following overall conclusions:

- As documented by the literature, a model-based technique for reconstructing traffic flows from sensor data is overall superior to a naïve, non-model-based one. In other words, for a given number of traffic detectors on a highway section, a model can provide more accurate information than simple interpolation would, though the gain can be very moderate;

- As the density of traffic monitoring stations increases nearby a critical value of ½ mile, the importance of location and data processing algorithms becomes negligible. While being data-smart is better than not, when it comes to deploying traffic sensors, budgets trump intelligence in a seemingly considerable ratio;

- For the most part, it looks like the best locations for traffic detectors are relatively stable with regards to the number of sensors for which those locations are optimized. This is good news, because practitioners can instrument a facility gradually without second thoughts, and because it provides a rationale for building redundancy at critical locations;

- Given their frequency, detector failures must be taken into account in the formulation of deployment guidelines. So far, some limited results show that doing so may have a significant impact on those guidelines, but more analysis is needed before any conclusion can be drawn.

While more work could be done to extract systematic results that exploit all of the data sets used in the project, the findings presented herein are sufficient to propose simple guidelines that can reassure practitioners about the validity of a given detector deployment. The effects of detector failures should be studied more in-depth, because it is a fact of life that a certain percentage of detectors are out of service at any given time. Finally, a more definitive set of metrics must be put forth to the industry and receive enough recognition to become authoritative.

In rural areas, sensor locations should be prioritized at the scale of a district or region. Table 1 lists location attributes and their relevance to a common set of ITS elements. Note that these attributes are not prioritized or weighted and this may be a direction for future research.

For dense urban areas where recurrent traffic congestion will tend to guide the need for traffic detection, the team developed the Sensor Allocation Program. This is an intuitive, Excel-based tool designed to suggest sensor locations along a freeway corridor. The program optimizes sensor placement to maximize the precision of a bottleneck detection algorithm, and estimates the expected accuracy of popular freeway performance measures such as travel times and delays. No fine-grained prior knowledge of traffic patterns is required. The model can operate from the same kind of demand forecast scenarios that serves planning purposes. In addition to suggesting an optimal sensor configuration, the Sensor Allocation Program can work off of existing detector locations to fill gaps, and also offer a repair priority schedule where failures are recorded. Based on the interest from Caltrans districts for this decision aid, the team can further improve its ease of use and its set of features.

Table 1: Location Selection Criteria

Location Criteria	RWIS	CCTV	Loops	DMS
Mountain passes	X	X		
Ski areas	X	X		
High wind, rain, snow, fog	X	X		
Common icy conditions	X	X		
Shaded areas	X	X		
Bridges	X	X		
High proportion of weather related crashes	X	X		
Flooding locations	X	X		
High frequency of road closures	X	X		
View of DMS		X		
High frequency of crashes		X		
Rest areas		X		
Major intersections and interchanges		X		
Structures		X		
Good view		X		
Straight road		X	X	X
Upstream of junctions				X
Upstream of chain up areas				X
MUTCD				X
Upstream of common weather events				X
Planned construction project	X	X	X	X
Available power	X	X	X	X
Available communication	X	X	X	X
Can visit from maintenance yard in one day	X	X	X	X
At maintenance yard	X	X	X	X
Not within 2 miles of same type of device	X	X	X	X
Co-located with other devices	X	X	X	X
Good access	X	X	X	X

REPORT CONTENT AND ORGANIZATION

This report is organized into six sections that correspond to the areas of investigations that the team performed over the course of the project.

The first section, titled **PROJECT SUMMARY**, provides a general overview of the entire project. It describes the objectives of the project, introduces key premises and assumptions, and articulates the overall methodology. It is concluded by a summary of findings, and it abundantly references other sections of the report where more details can be found.

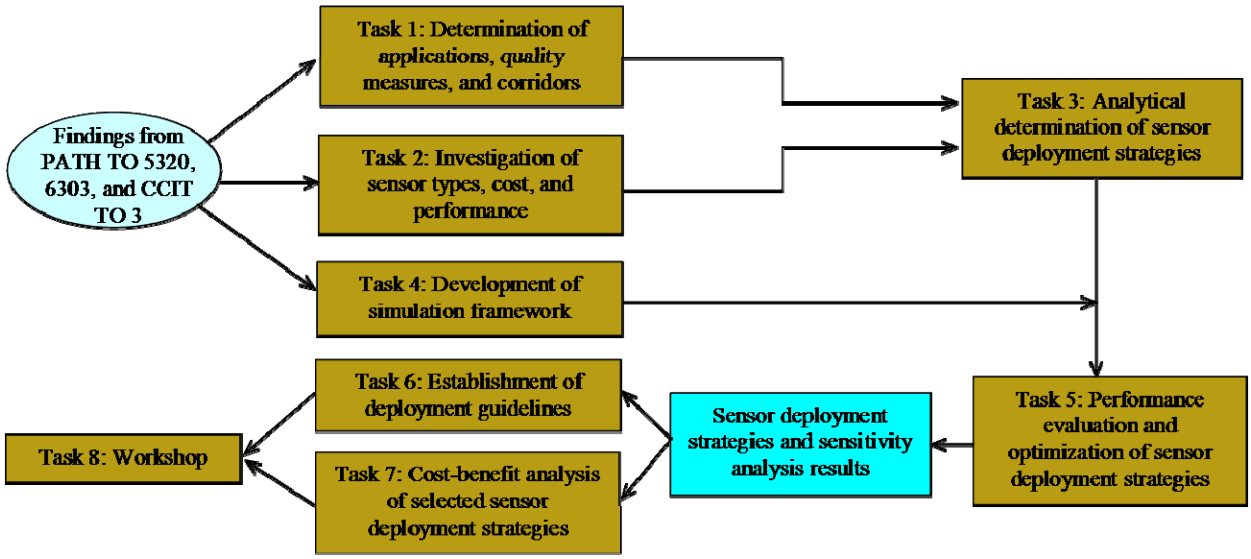


Figure 1 - Scope of Work

The second section, titled **PROJECT ASSUMPTIONS**, reports the results of tasks 1 and 2 of the scope of work, of which a schematic is provided for reference in Figure 1. It comprises a subsection on highway taxonomy, a subsection on modeling demand for a hypothetical highway corridor, a subsection on detector modeling and errors, and a subsection that describes the detector requirements of the ramp metering schemes in use in California.

The third section, **ANALYTICAL FRAMEWORK**, corresponds to tasks 3 and 4 of the scope of work. It lays the foundations of the simulation framework that was used for most of the project, and presents a purely analytical assessment of the effects of detector spacing on the accuracy of traffic estimations.

The bulk of the project consisted in developing several approaches and the corresponding tools in response to task 5. The results from this work have been split into two sections.

Section 4, **CORE STRATEGIES**, reports on the two main threads that were pursued under task 5. The first thread employs micro-traffic simulation models as the basic framework to match ground truth information against data collected from detectors. A dynamic programming algorithm solves the

problem of optimal placement when a fixed number of detectors are allocated to a given section of freeway. The second thread focuses on deriving the most utility from detectors by considering that consecutive detectors provide relevant information to the extent that they pickup variations in traffic conditions along a corridor. A benefits factor is developed to capture this idea, which is applied to a case study.

In section 5, we report on **ADVANCED STRATEGIES** that build upon the core strategies presented in section 4. The first subsection generalizes the dynamic programming algorithm to encompass the requirements of both accurate travel times estimations and freeway ramp metering. The second subsection offers refinements to the benefits factors method by proposing a more effective optimization scheme and leveraging it to introduce probabilities of detector failures as part of the search objective. The third subsection utilizes the micro-traffic simulation framework already employed to calibrate the dynamic programming method, but it introduces a more sophisticated treatment of the raw detector data. While most operators assign homogeneous traffic variable values to entire freeway segments that span consecutive sensors, the results presented in that third subsection assume that a traffic estimation scheme akin to the cell transmission model is put in use.

Section 6, **OPTIMAL DETECTOR PLACEMENT TOOLKIT**, presents a Microsoft Excel-based decision aid that is intended as a practical tool for transportation engineers and planners. In its current iteration, the tool implements the benefits factor method to recommend optimal detector locations based on planning data for a given freeway corridor. Based on the interest from Caltrans districts for this decision aid, the team can further improve its ease of use and its set of features. The development of the toolkit corresponds to task 6 in the scope of work. It is complemented by a set of guidelines aimed specifically at rural districts.

Note that task 7 (costs/benefits analysis) was not attempted because tackling it requires the formulation of a very large set of assumptions. This would not only have been costly, but also likely of limited value to Caltrans because of the ensuing controversy. We still believe that this task should be undertaken, but only after substantial coordination with Caltrans' Division of Traffic Operations and Districts has taken place, so that a dominant design can emerge to guide its objectives and methodology.

Task 8 is still pending but its completion is not required for the delivery of this report. CCIT will coordinate with Caltrans' Division of Traffic Operations to organize an information session and present the results of this project as well as the optimal detector placement toolkit to a group of transportation engineers and planners.

Following is a table of contents that summarizes the above information and provides page references.

Table of Contents

SECTION 1 – PROJECT SUMMARY

J.D. MARGULICI ET AL. “A FRAMEWORK FOR ANALYZING THE SENSITIVITY OF TRAFFIC DATA QUALITY TO SENSOR LOCATION AND SPACING”. ADAPTED FROM EPONYMOUS CONFERENCE PAPER, 15TH WORLD CONGRESS ON INTELLIGENT TRANSPORTATION SYSTEMS, 2008.

A FRAMEWORK FOR ANALYZING THE SENSITIVITY OF TRAFFIC DATA QUALITY TO SENSOR LOCATION AND SPACING

J.D. Margulici
Associate Director
California Center for Innovative Transportation
University of California, Berkeley
2105 Bancroft Way, Berkeley CA 94720-3830, United States
(510) 642-5929 – jd@calccit.org

Xuegang (Jeff) Ban, Rensselaer Polytechnic Institute, banx@rpi.edu
Alexandre Bayen, University of California, Berkeley, bayen@berkeley.edu
Lianyu Chu, University of California, Berkeley, lchu@calccit.org
Adam Danczyk, University of Minnesota, danc0010@umn.edu
Juan-Carlos Herrera, University of California, Berkeley, jcherrera@berkeley.edu
Ryan Herring, University of California, Berkeley, ryanherring@berkeley.edu
Henry X. Liu, University of Minnesota, henryliu@umn.edu
Olli-Pekka Tossavainen, University of California, Berkeley, optossav@berkeley.edu
Daniel Work, University of California, Berkeley, dbwork@berkeley.edu

INTRODUCTION

Various types of traffic detectors, including loop detectors, radars, toll tag readers and video cameras are deployed on highways. They provide the data needed to run traffic management applications such as ramp metering control, bottleneck identification, and travel times estimation. However, few studies have systematically analyzed the data requirements of these applications in terms of detector spacing and location, even if some have appeared in recent years (1), (2), (3), (4), (5), (6), (7). In other words, the trade-offs between the cost of detectors and their benefits for traffic estimation accuracy are not well known. As a result, most highway detectors are installed using ad hoc guidelines or on a case-by-case basis, rather than through the application of measurable objectives. This in turn makes it difficult for practitioners to justify equipment and maintenance expenditures, often slowing deployment.

This paper presents tools and results developed as part of the Partners for Advanced Transit and Highways' Task Oder 6328, sponsored by the California Department of Transportation (Caltrans). The objective of Task Order 6328 was to determine the sensitivity of traffic data quality to detectors location and spacing. Two factors are motivating this work. First, the development of traveler information fostered by the booming market for in-vehicle navigation and personal navigation devices (PND), as well as the growing emphasis by highway network operators on performance monitoring and management, are calling for a fresh look at how much data is needed. At the same time, cheaper alternatives to inductive loop detectors are now widely available and make it affordable to deploy dense networks of detectors. In order to guide deployment plans, a better sense of optimal sensor spacing is needed.

APPROACH AND TOOLS

As depicted in Figure 1, we considered three distinct steps involved in processing traffic data, from collection to assembling to utilization. A key premise of the project described in this paper is that data from traffic sensors is aggregated to provide information, and that the requirements

for that information must be driven by specific applications. While this may sound trivial, practical deployments are more often guided by what is feasible than by what is desirable. In fact, there exist no widely accepted quality targets or even standard metrics to describe the information requirements of typical freeway management applications such as travel times or delays estimation, bottleneck identification and ramp metering.

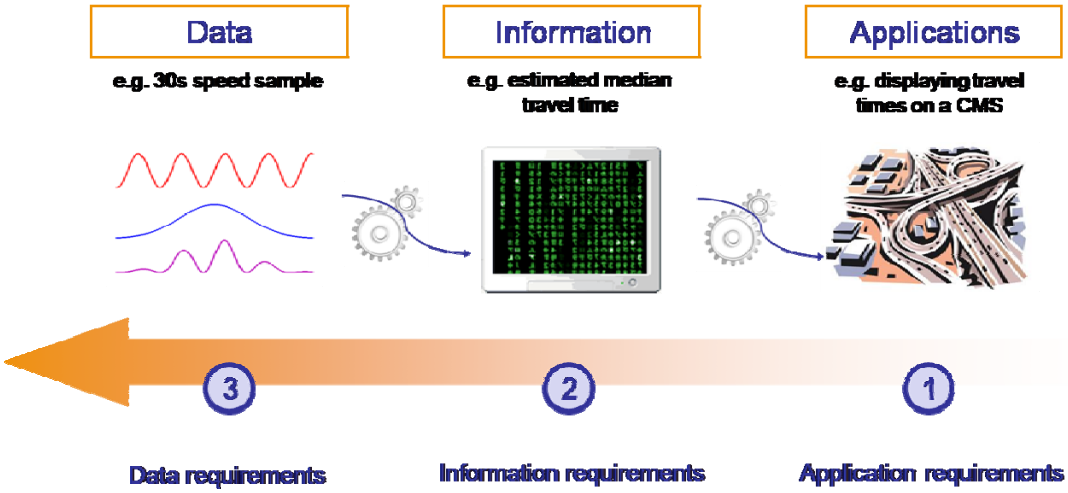


Figure 1 - Traffic Data Processing and Backward Flowing Requirements

Therefore, our first effort consisted in defining quality measures for those applications. Quality measures are then used to explore the relationship between sensor configurations on a corridor, which are defined as the mix of sensor types and locations, and the resulting performance of selected traffic applications. In other words, quality measures link sensor deployment strategies to application requirements.

Of course, in the real world, our knowledge of traffic conditions is usually provided by the very same sensors whose placement we would like to optimize. In order to study the outcome of a given sensor configuration in terms of information quality, we need data far more granular and reliable than that from sensors. Ideally, we would like to have perfect flow information, and quantitatively compare that information to what is obtained from traffic sensors. Such perfect information is virtually impossible to assemble, although it has been done as part of traffic experiments that are succinctly described in this paper and that we incorporated in our work. The alternative is to employ traffic simulation modeling. While such models are limited in their abilities to consistently reproduce actual traffic conditions, their limitations only carry a second-order effect in our study. This is because our focus is on comparing the information derived from a comprehensive set of trajectories generated by the model, with the information sampled from virtual sensors positioned at various locations on the roadway geometry. Even if the trajectories only mimic reality approximately, the relationship between the ground truth information they provide and the information obtained from the virtual sensors remain relatively unaffected unless the modeling assumptions are blatantly wrong.

Our general approach consists in using trajectory sets generated from field experiments and traffic simulation models as referent data, or ground truth. In order to treat the output from micro-traffic simulation, we implemented a traffic monitoring station (TMS) model in Matlab. This allows us to extract simulated sensor data from vehicle trajectory sets at arbitrary locations. This data is then processed to assemble meaningful information in the same way that a Transportation

Management Center (TMC) would. Note that we limited our analysis to point detectors such as loops or radars, leaving aside segment-based detection mechanisms such as toll tag readers. After running the Matlab model, the team may obtain, for instance, travel time estimates generated by sampled sensor data along with the actual travel times derived from vehicle trajectories. Information quality measures for selected applications are automatically generated, providing a score for the sensor configurations being tested. Another, Excel-based, tool was developed to run in conjunction with a more macroscopic freeway flow model. This tool focuses particularly on optimally positioning sensors to identify pockets of congestion.

Within this framework, we can simulate multiple corridors, employ several data processing algorithms, including simple instantaneous methods as well as model-based estimations, and leverage various optimization techniques to select sensor configurations that yield the best data quality scores. The paper describes the applications quality measures we selected, our experimental data sets, the traffic monitoring station model, the data processing techniques we simulated, and our search for optimal deployment solutions. An overview of results to date is outlined, and more detailed results can be found in other sections of the reports, which are pointed in the list of references.

APPLICATIONS OF TRAFFIC DATA AND QUALITY OBJECTIVES

A white paper prepared for the Federal Highway Administration (FHWA) recommended the following definition for traffic data quality (8): *Data quality is the fitness of data for all purposes that require it. Measuring data quality requires an understanding of all intended purposes for that data.* Traffic management applications have traditionally been grouped in three categories: traffic control, traveler information, and incident response. Those all require accurate, real-time traffic data in order to function optimally. Additionally, practitioners monitor traffic patterns and overall freeway performance for reporting and planning purposes. Hence freeway monitoring makes up a fourth application. This section describes the quality measures that we have identified for each application. The FHWA white paper suggests six generic measures of data quality, namely *Accuracy*, *Completeness*, *Validity*, *Timeliness*, *Coverage*, and *Accessibility*. This paper addresses issues of accuracy, validity and coverage. We essentially use root mean squared error (RMSE) metrics to compare ‘ground-truth’ values with estimated values of traffic variables.

Freeway monitoring

Freeway monitoring comprises traffic volume counts and vehicle classification, as well as the identification and quantification of recurring or non-recurring traffic congestion. Traffic congestion can essentially be characterized by locating bottlenecks and measuring the severity, extent and duration of the resulting slowdowns.

Traffic volumes and vehicle classification at a given location can be captured by a single sensor. Therefore, this study has no bearing on those applications. Practitioners will install sensors in-between major nodes of the road network in order to assemble meaningful volume and classification information. On the other hand, properly measuring traffic congestion requires an array of sensors. Here, for any given roadway section, there is a clear trade-off between the density of traffic detectors, which comes at a cost for installation and maintenance, and the resolution and reliability of information obtained by practitioners. For congestion monitoring, quality can be defined as the proximity between the estimation of select traffic variables or freeway performance metrics, and their true value. Common freeway performance metrics include vehicle-miles of travel (VMT) and vehicle-hours of travel (VHT). Average speed and travel times can also be employed, and present the advantage of being easily understood by the

traveling public. Finally, operators must be able to locate and measures bottlenecks in order to address to causes of traffic congestion. Our framework examines the following measures: travel times on select itineraries, speed contour maps, and bottleneck location.

Traffic control

Freeway traffic control applications may include access metering and variable speed limits. The former is deployed on a much wider basis in the US, especially California, and was the object of our study. There exist various algorithms that turn traffic detector data into ramp metering control instructions. Most algorithms that currently operate in practice do this locally. In that case, the algorithm is closely tied with the existence of detectors at predetermined locations, and the requirements for sensor location and spacing are provided by each algorithm (9), (10). In this sense, there is not much than can be optimized. In its most ideal version, ramp metering is adaptive and coordinated. That is, metering rates at each ramp evolve in a coordinated fashion to respond to current traffic conditions and maintain incoming flows so that freeway capacity remains at its peak. One example of such a scheme is the System-Wide Adaptive Ramp Metering (SWARM) method, developed by Delcan. Multiple versions of SWARM exist and have been tested by Caltrans. Each version corresponds to a sensor configuration that repeats itself at each ramp. While this systematic approach makes sense from an implementation standpoint, we were interested in determining whether there was an optimal way to install sensors for ramp metering detection needs. We did not consider the ultimate output of ramp metering, which would be a reduction in delays, because finding a relationship between the location of traffic detectors and the delay reduction enabled by ramp metering would be very elusive, given the complexities of interactions at hand. For instance, any sensor configuration could, by chance, produce a winning metering strategy. Rather, we had to think in terms of quality and quantity of relevant information available to a ramp metering algorithm. The premise of a system-wide algorithm suggests that knowing traffic conditions at all times and all locations, while not absolutely necessary, would ensure the most effective form of ramp metering. Thus, from an information standpoint, our objective was to monitor conditions on a freeway in a holistic manner. Our objective was therefore formulated in terms of a freeway density map. An accurate density map is arguably the most complete piece of traffic information one can think of to describe highway flow.

Traveler information

Real-time traffic information on a variety of media has gained popularity in urban areas where congestion and incidents frequently impact vehicle travel. Map applications typically display speed maps that are color-coded. Besides, accurate travel time estimates help commuters assess traffic, alleviate their stress, and make better route decisions. In many urban areas, travel time estimates are displayed on Dynamic Message Signs (DMS). We picked speed contour maps accuracy and travel times estimates as our quality objectives for traveler information applications, both of which we already selected as part of gauging freeway monitoring applications.

Incident response

In many locations, inductive loops have traditionally been installed to detect the formation of queues and identify incidents. In fact, the distance between sensors has sometimes been determined on that basis. However, the spread of cell phones have changed the reality of incident detection. Drivers frequently notify 911 when observing accidents or obstacles, and this constitutes a faster and more reliable channel than traffic detectors. Certain operators employ video cameras and automated incident detection, which suppress the need to resort to slowdowns as possible indices. For those reasons, we overlooked incident detection, which was deemed no longer relevant in the context of point traffic detector deployment.

Summary

Following is a list of the quality measures we ultimately selected to study the tradeoffs between sensor density and information quality. Each measure assumes that we hold ground truth information on the one hand (a strong assumption, but one that is discussed in the next section), and estimates from traffic sensors on the other hand.

Travel times estimation accuracy: the quality measure for travel time estimates along a corridor made up of K elementary segments, for a given traffic flow is $E = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K}(e_k^m)^2}{M}$, where e_k^m denotes the difference between the predicted travel time and the actual travel time of vehicle m over the k^{th} segment (11). The objective is to minimize this sum.

Speed maps and density maps estimation accuracy: The principle of a contour map is to discretize space and time in order to compute a single speed (or density) value for each (x,t) pair. Whether for speed or density, we build two maps: one from ground truth and one from traffic estimates produced by a given sensor configuration. We outline three methods for comparing the two maps. Each method produces a single metric that serves as a benchmark measure for evaluating sensor deployment strategies:

1. RMSE: differences between (x,t) pairs are squared and summed up to generate a single value from which we extract the square root. This is the most natural approach, but it can capture a lot of unimportant 'noise', i.e. small discrepancies between pairs that are not essential to estimation accuracy.
2. Alternatively, we only record differences that exceed a given threshold, e.g. 10 mph for speed and .025 for occupancy rates. This eliminates undesirable noise and only captures large errors.
3. The third method consists in discretizing the scale on which speed (or density) is represented. This is analogous to traffic information websites displaying 3-4 colors to indicate levels of congestion. The two maps can then be compared by calculating the surface area on which they match up perfectly, as a percentage of the total time-space area.

At this point, we are still experimenting with the respective merits of these methods, and setting the proper thresholds. Behind this investigation lies an important stake in the measurement of traffic data quality. Similar efforts are ongoing so that standards can be set industry-wide (12), (13), (14), (15).

Bottleneck detection accuracy: for bottleneck detection accuracy, we designed a 'benefit' factor that is based on speed gradients (16). The rationale is that the closer two traffic monitoring stations are placed upstream and downstream of a bottleneck, the more accurate the detection. Between those two TMS, a speed gradient will be observed when the bottleneck is active. An exact formulation of the corresponding quality objective can be found in (16), (17). The benefit factor between two consecutive TMS is essentially $b_{ij} = \frac{\sum_{t=0}^{T}(V_j^t - V_i^t)}{(S_j - S_i)}$, where a time interval is divided into T segments, V_i^t is the measured traffic speed at TMS i and time t, and S_i is the postmile of TMS i.

CORRIDOR SELECTION AND GROUND TRUTH TRAFFIC SAMPLES

In order to examine the relationship between sensor configuration and information quality, we have assembled a meaningful sample of data sets, ultimately aiming for generalization of our results. The basis for our methodology is to compare the information obtained from traffic sensors with so-called 'ground truth' data. Our ground truth data was provided by field experiments in two cases and by simulation models in others. To ensure that we would be looking at a representative set of corridors and scenarios, we first established a taxonomy (18). This led us to define rough criteria to discriminate between typical rural corridor characteristics and typical urban characteristics, as presented on Table 1.

Table 1 - Typical freeway corridor characteristics

	Urban	Rural
Number of lanes (per direction)	>3	2-3
Spacing between interchanges (miles)	0.5-1	>1
Free flow speed (mph)	60-70	70
Proportion of trucks	<5%	5-30%
Daily traffic volume (per lane)	15K-25K	5K-15K

Experimental data sets

For all of our technological capabilities, it is still very difficult to acquire high-resolution traffic flow on a transportation corridor without engaging considerable means. Two recent experiments in traffic data collection conducted through CCIT produced data sets that fulfilled our needs.

NGSIM

As part of FHWA's Next-Generation Simulation (NGSIM) program, CCIT worked with Cambridge Systematics to capture individual vehicle trajectories on a 1/3-mile stretch of Interstate 80 in Berkeley. Video data was taken at the Berkeley Highway Lab from bird eye view cameras, and processed by employing machine vision algorithms. The result is a complete description of the local traffic flow over periods of up to 45minutes (20).

Mobile Century

The Mobile Century field test took place in February, 2008, in the San Francisco Bay Area (21), (22). It was conceived as an experiment in traffic data collection from probe vehicles carrying GPS-equipped mobile phones. One hundred vehicles were deployed on a 10-mile stretch of Interstate 880 near Union City, Cal. This enabled us to reach a traffic penetration rate of close to 5%, which for all practical purposes can be assumed to represent ground truth. This is illustrated by Figure 2.

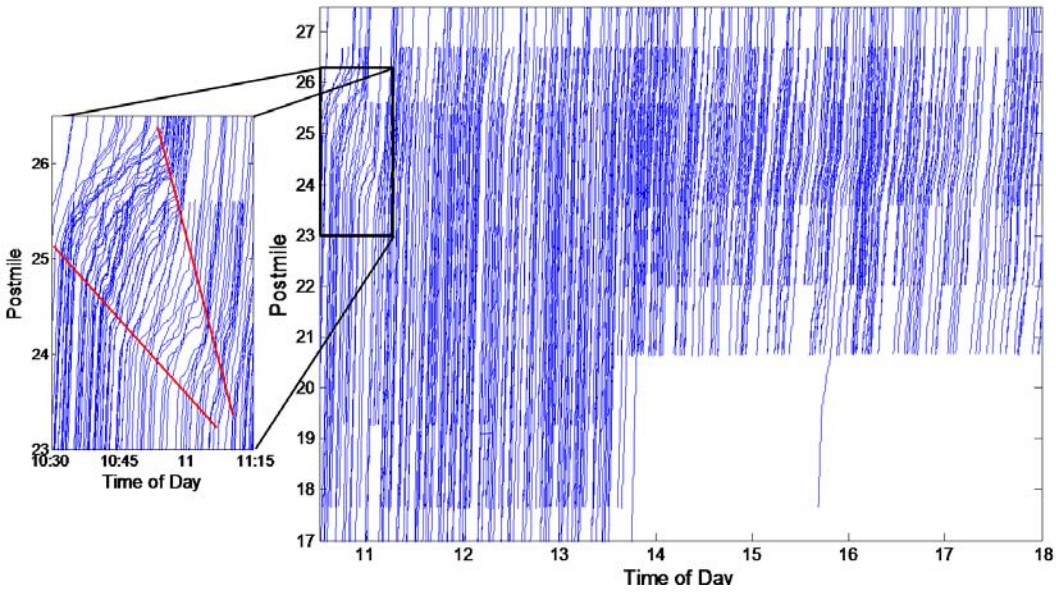


Figure 2 - Individual trajectories extracted from the Mobile Century experiment

Micro-simulation models

Case studies

Besides the two experimental sets, we employed simulation to construct realistic traffic flows, and then plugged in sensor models to recreate field conditions. In particular, we recycled existing micro-traffic simulation models developed with Paramics for the purpose of traffic operations planning (23). They included Interstate 880 (I-880) in the San Francisco Bay Area, and California State Route 41 (SR-41) near Fresno. The I-880 corridor corresponds to a heavily congested urban corridor. Baseline data is available as part of two models: one for the morning rush hour and one for the afternoon. For the purpose of this project, we selected a segment that is roughly 8.7 miles in length and ran the simulation for durations of roughly two hours (24). SR-41 traverses the Fresno urbanized area and southeastern Madera County. It serves as an example of a semi-urban corridor, and currently operates with moderate levels of congestion. The network is approximately 16 miles in length and the model includes two 3-hour periods in the AM and PM peaks (25).

Hypothetical corridors

We also developed tools to generate hypothetical corridors in Paramics. Our purpose here is to test the generalization of rules we may infer about the relationship between traffic detector configurations and information quality in a variety of settings. The hypothetical corridor model assumes that the freeway geometry is a straight line with repeating patterns, which can be automatically coded by a script, and the spacing in-between ramps can be varied. A traffic generator produces travel demand matrices on the corridor along the same principle: a basic pattern can be modified to recreate different congestion-inducing scenarios (26).

Cell transmission model

As an alternative to micro-traffic simulation models, which are very costly to develop, we also considered reproducing traffic conditions using the cell transmission model (CTM), a macroscopic approach. Initial traffic conditions can be assigned from field measurements if detector density is high. As is the case with micro-traffic simulation outputs, a detector model is then used to simulate the information collected from various sensor configurations. One of the benefits of this model is that it could be used as the final output of this project to recommend sensor configurations for corridors that are unequipped: traffic assignments could be pulled from planning and road design documents to constitute a baseline. A case study was conducted along Interstate 94 (I-94) in Minneapolis, Minnesota, a corridor that is 7.2 miles in length. Data was captured for entry and exit flows measured on that corridor from 16 existing loop stations during the PM peak rush hour on Wednesday, June 13, 2007 (17).

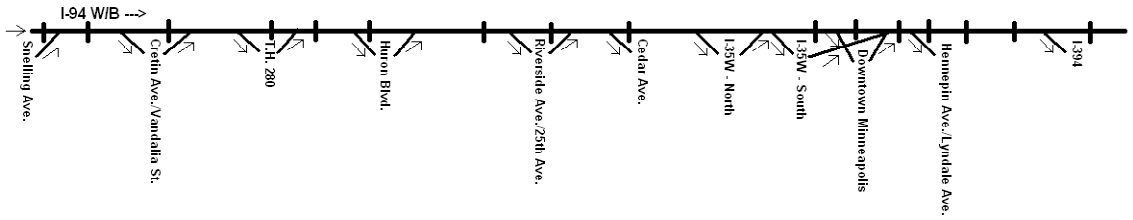


Figure 3 - I-94 layout

TRAFFIC MONITORING STATION MODEL

While inductive loops still dominate the industry, multiple technologies are available for point traffic detection. They include most notably magnetic, acoustic and radar (or infrared) detection, as well as image processing from video cameras (27), (28), (29). Many studies have been conducted to compare the respective merits of these technologies, in particular with respect to detection accuracy. However, technologies are evolving and studies often employ disparate methodologies, with sometimes contradictory conclusions. As a result, it is still difficult today to precisely assess the accuracy of concurrent technologies and products. In fact, the lead author has suggested that accuracy metrics and a methodology be developed through AASHTO in order to facilitate further testing of traffic detectors across the U.S. Nonetheless, and probably because of the difficulty in obtaining clear and consistent information in this area, we did not want to make this study about detection technology. Our goal is to examine the relationship between the number of traffic detectors and information quality on a corridor, the choice of technology being left to operators. Of course, this overlooks the fact that no traffic detector is perfectly accurate, and even further, that different technologies will cause different types of inaccuracies. Detector reliability and failure rate constitute another dimension that is even maybe more crucial. No freeway operator is able to maintain 100% functioning traffic detectors, and this means that information is lost. A robust sensor deployment plan must take this fact into account by making redundancy provisions and ensure that an acceptable level of information is satisfied at all times. This part of our work is still in progress: we have initially focused on developing data sets and tools to conduct this study, and we have conducted our present analysis by considering perfect detectors. However, introducing inaccuracies and failures is a logical next step that we have started to address.

Detector model

For the purpose of this study, detector stations are modeled as fixed probes in a given traffic flow, recording elementary traffic variables over a specified time interval, typically 30 seconds. Those traffic variables include volume, occupancy rate and average speed. Thus, each station behaves as a traditional TMS, such that would be enabled by inductive loops. We do not consider individual detection events or lane-by-lane data, although, in the case of micro-simulation, this information capture is emulated in the background. The model's principle is illustrated in Figure 4.

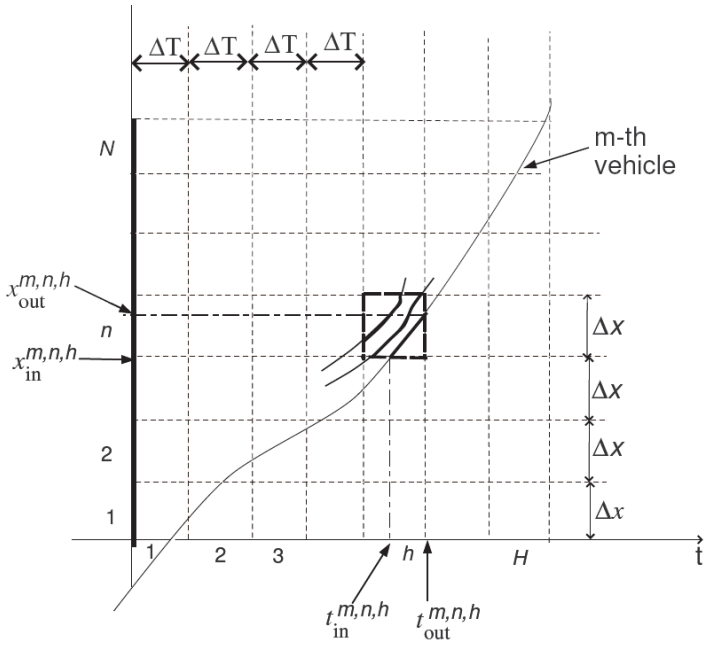


Figure 4 - Principle of TMS model in the case of a micro-simulation or traffic experiment output

There is a strong consensus across the ITS industry to recognize that well-functioning inductive loops are a trustworthy instrument. Their main disadvantage is that they must be installed within pavement at the expense of lane closures and costly maintenance. On the other hand, non-intrusive detectors such as radars have not always shown the same reliability as loops. In order to account for inaccuracies, we make the simplifying assumption that operators can either choose expensive, intrusive detectors that are relatively error-free, or non-intrusive detectors that are cheaper but slightly inaccurate. Therefore, one version of the detector model outputs the exact traffic variables derived from flow data, while the other version adds an error term to each of them.

Detector inaccuracies

To model the errors made by non-intrusive sensors, we used former studies dealing with sensors errors (30), (31), (32), (33), (34), (35), (36). The results given by those studies were extremely disparate but presented a few commonalities which led us to two key design decisions. First, the error term is bi-modal with respect to traffic conditions. Almost all studies agree on the fact that sensor data is less accurate during congestion conditions than during free-flow. We chose a single velocity criterion (45 mph) as the limit between congestion and free-flow. The second feature of the model is that errors are normally distributed. However, the speed error model attributes an absolute error expressed in miles-per-hour, while the count and occupancy error models attribute a relative error in percents. This is because the latter variables scale up and down with traffic,

while the former does not. Further, we assumed that the errors on count or occupancy would be identical, while the error on speed would be a second, independent term. Note that for very low speeds, negative values may occur after accounting for errors. When that is the case, the error term is redrawn until the corrected speed is a positive value. Thus, the complete error model can be formulated as follows:

$$\hat{q}(x_0,t) = \bar{q}(x_0,t).(1+\tilde{q}.e_1(x_0,t)) \quad (1)$$

$$\hat{k}(x_0,t) = \bar{k}(x_0,t).(1+\tilde{k}.e_1(x_0,t)) \quad (2)$$

$$\hat{v}(x_0,t) = \bar{v}(x_0,t) + \tilde{s}.e_2(x_0,t) \quad (3)$$

Where $\hat{q},\hat{k},\hat{v}$ designate volume, occupancy rate and speed estimates at location x_0 for time interval t; $\bar{q},\bar{k},\bar{v}$ designate true time-average values; $\tilde{q},\tilde{k},\tilde{s}$ are constant, and e_1, e_2 are standard normal distributions. In free flow conditions and for time intervals of *5 minutes*, a compilation of past studies of various technologies led us to select the following values:

Free flow conditions: $\tilde{q} = 6\%$ $\tilde{k} = 6\%$ $\tilde{s} = 1.6\ mph$

Congested conditions: $\tilde{q} = 8\%$ $\tilde{k} = 8\%$ $\tilde{s} = 4.8\ mph$

These numbers must be scaled up or down if different time intervals are selected.

Detector failures

Detector failures are relatively easier to simulate if we simply assume that failures result in missing data. We attribute a probability of failure to each traffic monitoring station, and a failed station is like no station. In effect, accounting for detector failures will impact both the recommended density and locations of traffic monitoring stations. First, more sensors are needed in order to maintain the level of quality that could be reached without failures. Second, optimal solutions will feature redundant sensors at locations that are critical from an information standpoint. Other solutions would receive lesser scores because a failure at a critical location end up being penalizing. Integrating this dimension into sensor deployment planning will represent an improvement over current practice. An attempt to integrate detector failures into the planning of optimal detector placement is presented in Section 5, *Advanced Strategies* (17).

DATA PROCESSING TECHNIQUES AND ESTIMATION ALGORITHMS

The link between raw detector data and traffic information is far from obvious and has been examined through hundreds of studies that propose or evaluate various processing algorithm, ranging from clever refinements over tried techniques to highly sophisticated statistical tools. A paper presented at the 14th ITS World Congress inventories no less than 40 different approaches to travel time estimates (26). From that standpoint, it is not practically possible to establish an absolute relationship between sensor data and information quality, because the latter wholly depends upon processing techniques. Yet, a careful examination of all possible techniques reveals that there is no silver bullet: abundant data results in good quality estimates while the most refined algorithm will stall from lack of data. From this, we contend that variations in algorithms have a second-order effect at best. Nonetheless, we still draw a fundamental distinction between model-based approaches, which assume some underlying knowledge of typical traffic flow properties, and non-model-based, or ‘naïve’ approaches.

Naïve methods

What we call the naïve approach consists in dividing up a roadway segment into independent cells that each correspond to a single traffic monitoring station. For each TMS, the corresponding cell is also called 'area of influence.' A single value is attributed to each traffic variable in a given cell over time, based on a direct averaging of the data provided by the detectors that make up the TMS. Notwithstanding its naïve qualification, and in spite of abundant research literature condemning it, the naïve method is probably used, give or take, by at least 90% of traffic operators and traveler information systems in the world. Therefore, for all this method's shortcomings, it is most important to resort to it to study the effects of detector density on information quality. In fact, as detector density increases, the naïve method produces good results, and it has the advantage of providing very reliable, even if biased, outputs. By contrast, fancier methods may fare better with less data, but they are also prone to dubious outputs, especially when they behave like 'black boxes.' With traffic detection costs falling, the naïve method still has a bright future, and this justifies the value of the present study.

Kalman filtering

In order to contrast the use of the naïve approach with a model-based approach, we selected an improvement upon the cell transmission model as a proven method for estimating flow variables from sparse data (38). In the CTM, roadways are also divided up into cells, but those can be much smaller in length than the distance between TMS, and it may therefore be that only a few of them actually contain traffic detectors. CTM employs flow conservation equations and Kalman filtering to estimate traffic variables in each cell. Note that CTM is used here as an estimation method. This is completely independent from its use as a traffic simulation technique on Interstate 94, though the underlying principles are of course the same.

Bottleneck identification from neighboring sensors

In the case of bottleneck detection accuracy, there is congruence between the quality objective and the data processing technique. A benefit factor is defined by calculating speed gradients in-between TMS. Though it has no direct application to traffic engineering in and out of itself, the benefit factor is derived from raw traffic detector data (16).

OPTIMIZING SENSOR LOCATION

Having defined information quality measures, produced set of realistic traffic flows, and implemented data capture and processing tools, we can work on optimizing the location of traffic detectors. The search space is made up of arrays of freeway postmiles describing the positions of traffic monitoring stations. The objective function is one of the quality measures described in the corresponding section of this paper. Search is conducted after setting the number of TMS, and we vary this number in order to find an optimum for any 'budget.' Various optimization techniques are used, depending on the formulation of the problem, which is itself based on both the objective function and the algorithm used to process the data.

Dynamic programming formulation

In the case of the naïve data processing method, we determined that the optimal sensor location problem for travel time estimation can be formulated as a Dynamic Programming (DP) model (11), (38). This formulation also applies to the optimization of quality objectives based on speed or density contour maps. The search can be done in polynomial time, and we implemented it to

run in batch mode for a range of number of TMS. We also showed that incorporating sensors that have already been deployed on a roadway segment and finding optimal locations to fill the gaps can be done with the same formulation.

Linearization of the benefit factor optimization problem

When benefit factors are used to optimize the identification of bottlenecks by traffic detectors, a difficulty arises in that the problem formulation is non-linear. A linearized optimization model for aggregated-point sensors was proposed, based on a previous nonlinear model solving the same problem (17). This linearized model is far superior to its nonlinear counterpart, as it can use a traditional solver or a resource-constrained shortest path algorithm to find the optimal solution.

Random sampling

When the CTM model is used for estimating traffic variables, the optimization problem cannot be formulated analytically, regardless of which objective function is chosen. In this case, the search for an optimum is necessarily blind and must rely on non-traditional methods. On the other hand, the search space is reasonably small and the variations between different sensor configurations are of limited magnitude. Based on these observations, we simply considered randomly generated sets of sensor configurations in order to look for an approximate optimum in the case of Kalman filtering estimation. The random generation of the sensor configuration sets was constrained enough that only realistic configurations were considered.

RESULTS AND ENSUING GUIDELINES

This section presents the results that have been obtained to date from the framework described until this point. Detailed results are available in the referenced papers that have been produced by the project team. We offer both quantitative and qualitative results, but our goal is to turn those results into more generalized rules of thumb that can readily be used by practitioners, realizing that there is usually little knowledge available about non-instrumented corridors.

As the final product of this project, we propose an Excel-based Sensor Allocation Program that can recommend a sensor deployment strategy for a given corridor based on its geometry, the locations of the on-ramps and off-ramps, known demand patterns, as well as contextual information provided by practitioners, such as recurring congestion. This toolkit is provided in the report, though it is still work in progress and requires inputs from practitioners.

Influence of data processing techniques

As we expected, the model-based approach to traffic variable estimations produces better results than the naïve approach. This can be readily illustrated by Figure 5, which compares the quality of travel time estimates obtained from the cell transmission model and from the instantaneous method with various numbers of sensors on the I-880 study corridor. While this is not illustrated here, the optimal location of traffic detectors is also different depending on which estimation method is used. The naïve method essentially performs averages and does not ‘guess’ how traffic evolves. Optimizing for the naïve method leads to sensor configurations that cover the roadway section of interest as extensively as possible. By contrast, a model-based method is more ‘information-hungry’ and looks for singularities in the traffic flow, while segments that are mostly free-flowing do not need as much coverage. Similarly, the use of benefit factors to optimally place sensors results, by design, in configurations where TMS are positioned immediately downstream and gradually upstream of known bottlenecks.

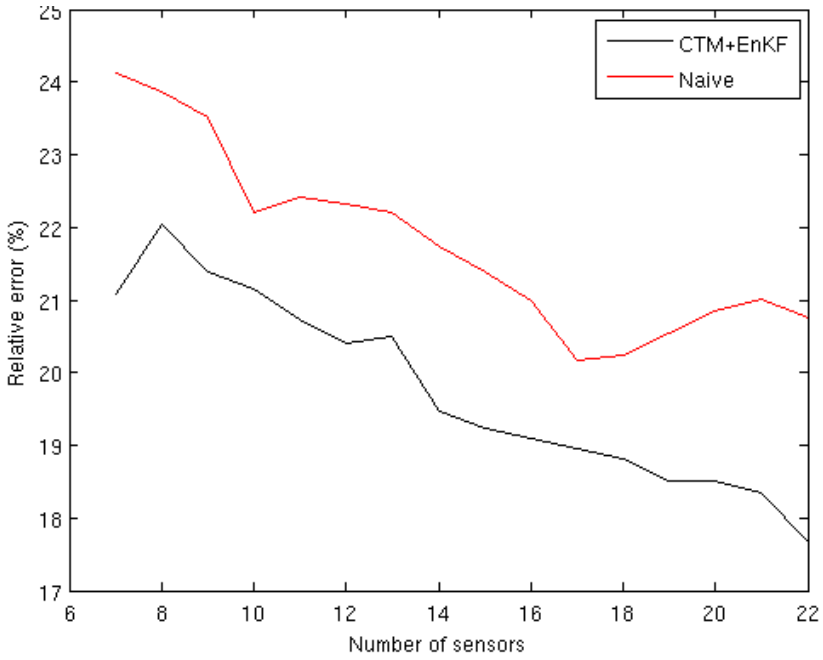


Figure 5 - Comparison of CTM-based and naïve estimation for travel times (I-880 case study)

As the density of sensors increases, the difference in configurations obtained with each method will not be as pronounced, if only because the available space merely becomes crowded! This leads us to the most essential consideration of this study, i.e. the overall relationships between sensor density and data quality.

Relationship between sensor density and data quality

Regardless of which estimation method is employed, information quality goes up as the number of traffic monitoring stations per unit of roadway length increases (18), (1), (2)**Error! Reference source not found.**, (4). Figure 6 exhibits the relationship between sensor spacing and the average error on a speed contour map in the case of the I-880 study corridor, using the naïve estimation method. The curves show a range of random configurations, the evenly spaced configuration and the optimal configuration, respectively. Consistent with previous studies, we notice that a ½ mile spacing in-between TMS produces errors less than 4 mph on average, which is acceptable if we consider that this corresponds roughly to a 10% error (in reality, the error is not distributed evenly in space and time, but this at least provides an order of magnitude.) Below ½ mile, the error continues to improve, but significantly more detectors are required to decrease spacing, which comes at a substantial cost (8). Another notable result is that the variance due to specific locations decreases as the density increases. This is consistent with the observation formulated in the previous section: once the density of data becomes large enough, the information improves. This holds regardless of exact detector locations or estimation techniques.

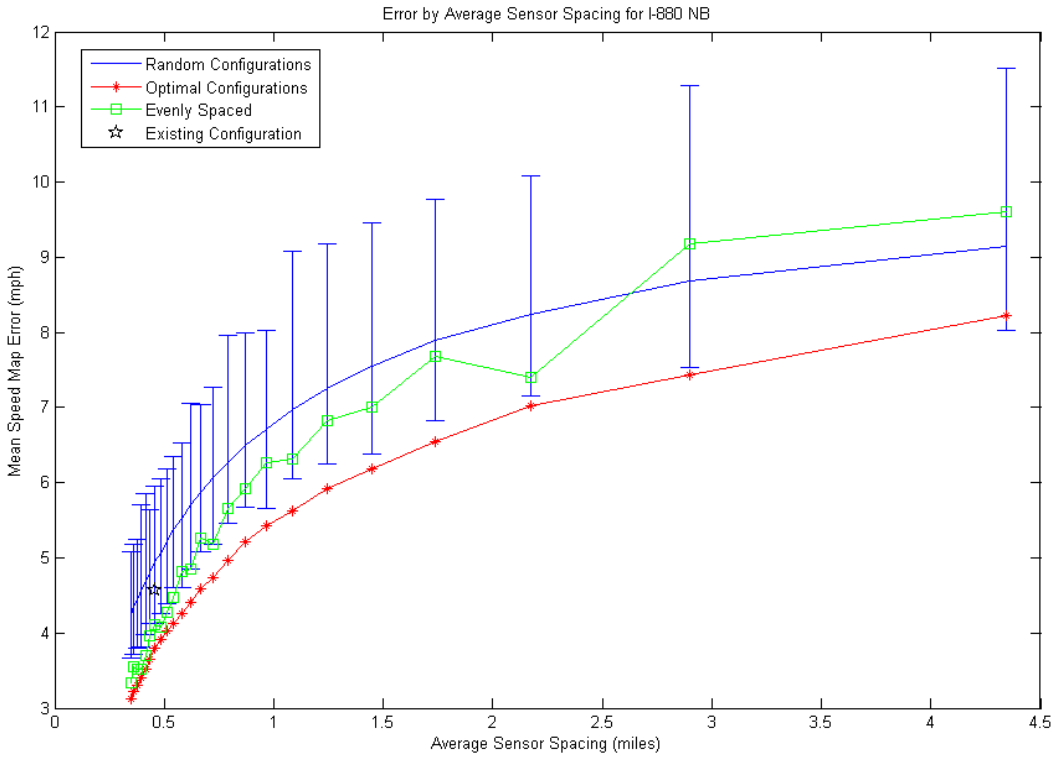


Figure 6 - Relationship between sensor spacing and information quality (I-880 case study)

Stability of preferred sensor locations

Another interesting question is the extent to which certain locations matter particularly to ensure information quality, regardless of the number of sensors deployed. If optimal locations shift as the number of sensors changes, then there is a dilemma. First, practitioners may want to start instrumenting corridors at a certain level and then add additional sensors as budgets become available. If the optimal locations for 8 TMS are different from the ones for 12 TMS, then there is no satisfying incremental deployment path. This problem is tempered by the fact that location is a less important factor as density increases. Second, practitioners must also deal with the very common occurrence of detector failures. If the optimal locations for 6 or 7 TMS are different than those for 8TMS, then a compromise might be preferable, because actual operations may often run with degraded instruments. Likewise, knowing detector locations that are always interesting is a good guide for building redundancy into the detection system. One caveat with this analysis is that it assumes that links can be dynamically reconfigured based on detector availability. In reality, most systems will simply show missing data on links for which the corresponding TMS is down. Figure 7 shows the set of optimal locations for increasing numbers of detectors in the case of the I-880 study corridor, using the naïve approach to estimate speed. The chart exhibits both sensor locations and link extremity locations, thus illustrating the aforementioned caveat.

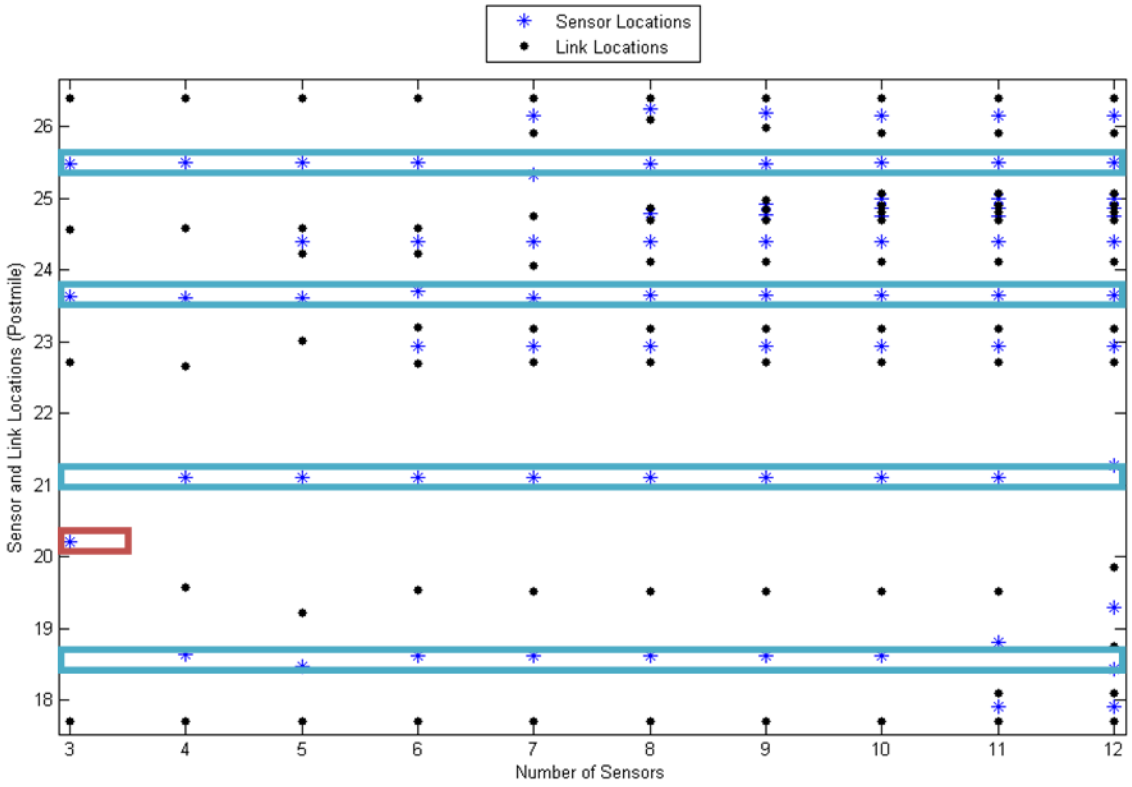


Figure 7 - Stability or instability of optimal sensor locations as density increases

Inclusion of inaccuracies and failures

Given the rate of detector failures observed in the field, no analysis of optimal sensor deployment strategies can be complete without including this dimension. Detection inaccuracies must also have an influence on the actual relationship between detector density and information quality, i.e. displacing the curve, but the effect is probably moderate with reasonably accurate sensors. We have not studied it so far. However, we have combined the possibility of detector failure with the search for optimum locations to identify bottlenecks. This was done by attributing a functioning probability of 80% to each TMS, which turns the optimization problem into an uncertainty-bound problem. In order to linearize the problem, it was assumed as a first step that only one TMS could fail at a time. Preliminary results are rather surprising, showing significant discrepancies between optimal configurations with or without the failure hypothesis (17). We plan on further investigating this topic in order to formulate guidelines for practitioners.

CONCLUSION

Our objective is to formulate detector deployment guidelines that are based on functional needs. Previous work has emphasized the design of algorithms that make the best use of available data, but less often questioned objective data requirements in the first place. We present a framework to study the relationship between sensor configurations and the quality of corresponding traffic variable estimates. This framework incorporates the following elements:

- Quality measures for selected traffic applications, so that information quality can be objectively quantified and various sensor configurations can be benchmarked accordingly;
- An extensive range of study corridors for which highly detailed flow data is available, either through simulation or from cutting-edge traffic experiments. Hypothetical corridors and scenarios can be automatically generated into Paramics models;
- A sensor model to simulate data capture for a given traffic flow, including a model of detection inaccuracies;
- Tools to process data generated by the sensor model according to various traffic estimation algorithms, most notably a naïve method and a CTM-based method;
- Optimization techniques and tools to rapidly identify the best sensor configuration for a given traffic flow, objective function and data processing method;
- An integrated, Excel-based tool that can be customized to accommodate future corridors and constitute a practitioner-friendly decision aid.

Results obtained so far draw the following conclusions:

- As documented by the literature, a model-based estimation technique is overall superior to a non-model-based one, though the gain can be very moderate;
- As the density of traffic monitoring stations increases nearby a critical value of ½ mile, the importance of location and data processing algorithms becomes negligible. While being data-smart is better than not, when it comes to deploying traffic sensors, budgets trump intelligence in a seemingly considerable ratio;
- For the most part, it looks like the best locations for traffic detectors are relatively stable with regards to the number of sensors for which those locations are optimized. This is good news, because practitioners can instrument a facility gradually without second thoughts, and because it provides a rationale for building redundancy at critical locations;
- Given their frequency, detector failures must be taken into account in the formulation of deployment guidelines. So far, some limited results show that doing so may have a significant impact on those guidelines, but more analysis is needed before any conclusion can be drawn.

While more work could be done to extract systematic results that exploit all of the data sets presented in this paper, the findings presented herein are sufficient to propose simple guidelines that can reassure practitioners about the validity of a given detector deployment. The effect of detector failure should be studied more in-depth, because it is a fact of life that a certain percentage of detectors are out of service at any given time. Finally, a more definitive set of metrics must be put forth to the industry and receive enough recognition to become authoritative. An important byproduct of this work will be the possibility to quantify, even roughly, the benefit of additional detectors in terms of traffic information accuracy. Being able to boast quantifiable benefits will bode well with the need to justify ITS budgets.

ACKNOWLEDGMENTS AND REFERENCES

The project team thanks the California Department of Transportation for its sponsorship and support.

References that point to other sections of this report are highlighted in bold characters.

(1) S. Jung, A. Toppen, K. Wunderlich. “The Effect of Average Loop Detector Spacing on the Accuracy of Calculated Travel Times: Twin Cities Case Study”. Mitretek Systems, 2005.
(2) I. Fujito, R. Margiotta, W. Huang, W.A. Perez. “The effect of sensor spacing on performance measure calculations”. In Proceedings of the 85th Annual Meeting of TransportationResearch Board, 2006.
(3) S. Oh, B. Ran, K. Choi. “Optimal detector location for estimating link travel time speed in urban arterial roads”. In Proceedings of the 82nd Annual Meetings of the Transportation Research Board, 2003.
(4) S.M. Eisenman, X. Fei, X. Zhou, H.S. Mahmassani. “Number and location of sensors for real-time network traffic estimation and prediction: A sensitivity analysis”. Transportation Research Record, 1981:253-259, 2006.
(5) H.D. Sherali, J. Desai, H. Rakha. “A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times”. Transportation Research B, 40:857-871, 2006.
(6) B. Bartin, K. Ozbay, C. Iyigun. “A clustering based methodology for determining the optimal roadway configuration of detectors for travel time estimation”. Transportation Research Record, 2000:98-105, 2007.
(7) P. Edara, J. Guo, B. Smith, C. McGhee. “Optimal Placement of Point Detectors on Virginia’s Freeways: Case Studies of Northern Virginia and Richmond”. Virginia Transport Research Council Report 08-CR3, 2008.
(8) Federal Highway Administration. “Defining and Measuring Traffic Data Quality”. White paper, Office of Policy, 2002.
(9) L. Chu. “Detector requirements for ramp metering in California”. California Center for Innovative Transportation, 2008.
(10) X. Ban, L. Chu, R. Herring, JD Margulici. “Optimal Sensor Placement Considering both Traffic Control and Traveler Information”. Submitted to 88th Annual Transportation Research Board Meeting, 2008.
(11) X. Ban, R. Herring, JD Margulici, A. Bayen. “Optimal Sensor Placement for Freeway Travel Time Estimation”. Submitted to the International Symposium on Transportation and Traffic Theory, 2008.
(12) F. Dance, D. Gawley, R. Hein, R. Kates. “Enhancing Navigation Systems with Quality Controlled Traffic Data”. Society of Automotive Engineers, Paper Number 2008-01-0200, 2007.
(13) J.D. Margulici, X. Ban. “Benchmarking Travel Time Estimates”. Conference paper, 14th World Congress on Intelligent Transportation Systems, 2007. Pending publication in IET Intelligent Transportation Systems.
(14) K. Ahn, H. Rakha, D. Hill. “Data Quality White Paper”. Center for Sustainable Mobility at the Virginia Tech Transportation Institute, 2008.
(15) S. Turner, B. Smith, M. Fontaine. “Developing a Standard Test Procedure for Travel Time Data Quality Assessment”. Scope of pooled fund solicitation #1206. Retrieved from http://www.pooledfund.org, 2008.
(16) H. Liu, A. Danczyk. “Optimal Detector Placement for Freeway Bottleneck Identification”. Conference Paper, 87th Annual Transportation Research Board Meeting, 2007.
(17) H. Liu, A. Danczyk. “An Integer Linear Program for Optimizing Sensor Locations along Corridors”. Submitted to Transportation Research, Part B, 2008.
(18) J. Kwon, B. McCullough, K. Petty, P. Varaiya. “Evaluation of PeMS to improve the congestion monitoring program”. Partners for Advanced Transit and Highways, 2006.
(19) J.C. Herrera. “A Taxonomy of Freeway Corridors”. California Center for Innovative Transportation, 2007.
(20) http://ngsim.camsys.com/

(21) J.C. Herrera et al. “Mobile Century, Using GPS Mobile Phones as Traffic Sensors: A Field Experiment”. Submitted to 15th World Congress on Intelligent Transport Systems, 2008.
(22) http://traffic.berkeley.edu/
(23) California Center for Innovative Transportation. “Corridor Management Plan Demonstration”. Final Report, Task Order 3, 2007.
(24) X. Ban, L. Chu, H. Benouar. “Bottleneck Identification and Calibration for Corridor Management Planning”. Transportation Research Record, 1999:40-53, 2007.
(25) H. Liu, S. Jabari, S. Hague, C. Castaneda. “California SR41 Corridor Simulation Study, Draft Final Report”. California Center for Innovative Transportation, 2007.
(26) J.C. Herrera. “Demand Generation”. California Center for Innovative Transportation, 2008.
(27) L. Klein, M. Mills, D. Gibson. “Traffic Detector Handbook: Third Edition—Volume I”. Turner-Fairbank Highway Research Center, FHWA-HRT-06-108, 2006.
(28) L. Klein, M. Mills, D. Gibson. “Traffic Detector Handbook: Third Edition—Volume II”. Turner-Fairbank Highway Research Center, FHWA-HRT-06-139, 2006.
(29) L.E. Mimbela, L. Klein. “A Summary of Vehicle Detection and Surveillance Technologies used in Intelligent Transportation Systems”. The Vehicle Detector Clearinghouse, 2007. http://www.nmsu.edu/~traffic/
(30) F. Herve, P. Supawanich. “Generic Error Model for Traffic Detectors”. California Center for Innovative Transportation, 2007.
(31) P. Martin, Y. Feng, X. Wang. “Detector Technology Evaluation”. University of Utah, MPC Report No. 03-154, 2003.
(32) J. Bonneson, M. Abbas. “Video Detection for Intersection and Interchange Control”. Texas Transportation Institute, The Texas A&M University System, Report 4285-1, 2002.
(33) J.D. Margulici et al. “Evaluation of Wireless Traffic Sensors by Sensys Networks, Inc.” California Center for Innovative Transportation, 2006.
(34) Pennsylvania Department of Transportation. “Traffic Data Collection Methodologies”. FHWA-PA-2006-005-040219, 2006.
(35) B. Coifman. “An Assessment of Loop Detector and RTMS Performance”. Partners for Advanced Transit and Highways, UCB-ITS-PRR-2004-30, 2004.
(36) Minnesota Department of Transportation. “Evaluation of Non-intrusive Technologies for Traffic Detection”. 2002.
(37) C.P.IJ. van Hinsbergen, J.W.C. van Lint, F.M. Sanders. “Short Term Traffic Prediction Models”. Conference paper, 14th World Congress on Intelligent Transportation Systems, 2007.
(38) D. Work, O.P. Tossavainen, S. Blandin, A. Bayen, T. Iwuchukwu, K. Tracton. “An ensemble Kalman filtering approach to highway traffic estimation using GPS-enabled mobile devices”. Submitted to the 47th IEEE Conference on Decision and Control, 2008.

SECTION 2 – PROJECT ASSUMPTIONS

J.C. HERRERA. “A TAXONOMY OF FREEWAY CORRIDORS”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2007.

J.C. HERRERA. “DEMAND GENERATION”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2008.

F. HERVE, P. SUPAWANICH. “GENERIC ERROR MODEL FOR TRAFFIC DETECTORS”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2007.

L. CHU. “DETECTOR REQUIREMENTS FOR RAMP METERING IN CALIFORNIA”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2008.

Corridor taxonomy

Geometric and operational characteristics are expected to vary from urban to rural corridors. By identifying these geometric and traffic attributes of urban and rural corridors, hypothetical corridors can be developed. Therefore, the purpose of this investigation is to identify common characteristics of urban and rural corridors. These hypothetical corridors will be used to analyze sensor requirements. It is important to note that the features identified in this study are *typical* features on urban or rural corridors. They do not intent to identify common features to *all* the urban or rural facilities.

Urban freeway traffic data is based on sample urban facilities I-280 N in San Francisco and US-50 E in Sacramento. The rural freeway traffic data is based on statistics from sample rural facilities I-80 E in Truckee, and SR 152-W in Merced. Information collected from PeMS[1] at these locations was used in this study. Sample size was limited by the lack of VDS on rural facilities.

Two geometric features (number of lanes per direction and spacing between interchanges), and three operational attributes (free flow speed, proportion of trucks, and daily traffic volume per lane) were chosen. It is important to note that congestion and delay were also investigated but they turned out to be specific to each facility and trends were not observed to differentiate incidence of congestion on urban or rural facilities. Table 1 summarizes the results found.

Table 1: Typical geometric and operational attributes of urban and rural corridors.

	Urban	Rural
Number of lanes (per direction)	$\geqslant 3$	2 - 3
Spacing between interchanges (miles)	0.5 - 1	> 1
Free flow speed (mph)	60-70	70
Proportion of trucks	$< 5\%$	5-30%
Daily traffic volume (per lane)	15K-25K	5K-15K

Urban corridors generally have three or more lanes in each direction of travel, and the interchange density is typically between one and two interchanges per mile. Free flow speeds on urban facilities are generally about 60 to 70 miles per hour.

Rural corridors generally have two to three lanes in each direction of travel. Interchange density on these facilities is typically less than one interchange per mile. Free flow speeds on rural facilities are generally 70 miles per hour or greater.

[1] `http://pems.eecs.berkeley.edu/`

The proportion of trucks varies during the day, specially for the rural case where truck volumes are typically greater than 5% of all traffic, up to 30%. On urban corridors, truck volumes are typically less than 5% of all traffic. In both the urban and the rural cases, the pattern of the proportion of trucks is almost the same during weekdays and during weekends. Figure 1 sketches how the proportion of trucks typically changes during the day at urban and rural corridors.

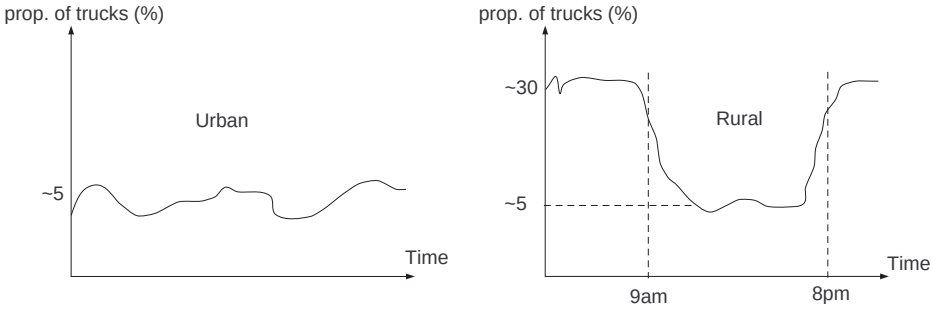


Figure 1: Proportion of trucks.

The traffic pattern not only differs from urban to rural corridors, but it also differs from weekdays to weekends. Figure 2 sketches typical traffic profiles for urban and rural corridors during weekdays and weekends.

Urban corridors: Daily traffic volumes are on the order of 15,000 to 25,000 vehicles per lane per day, with greater volumes occurring on weekdays than on weekends. During weekdays, flow levels increase between 5am and 6-7pm, with a marked peak during either the morning (as in Figure 2) or the evening. The flow level remains almost stable during the rest of the day. During weekends, flow increase starts later than during weekdays (9-10am). The level of flow remains almost constant during the day.

Rural corridors: Traffic characteristics may exhibit high levels of seasonal variation in this case. Daily traffic volumes are on the order of 5,000 to 15,000 vehicles per lane per day, with slightly greater volumes occurring on weekends than on weekdays. During weekdays, flow levels increase between 6am and 6-7pm. The flow level remains almost stable during this period. A peak period could eventually be observed at some facilities during either the morning (as in Figure 2) or the evening. During weekends, flow increase starts later and finishes earlier than during weekdays (9-10am to 4-6pm). The level of flow remains almost constant during the day.

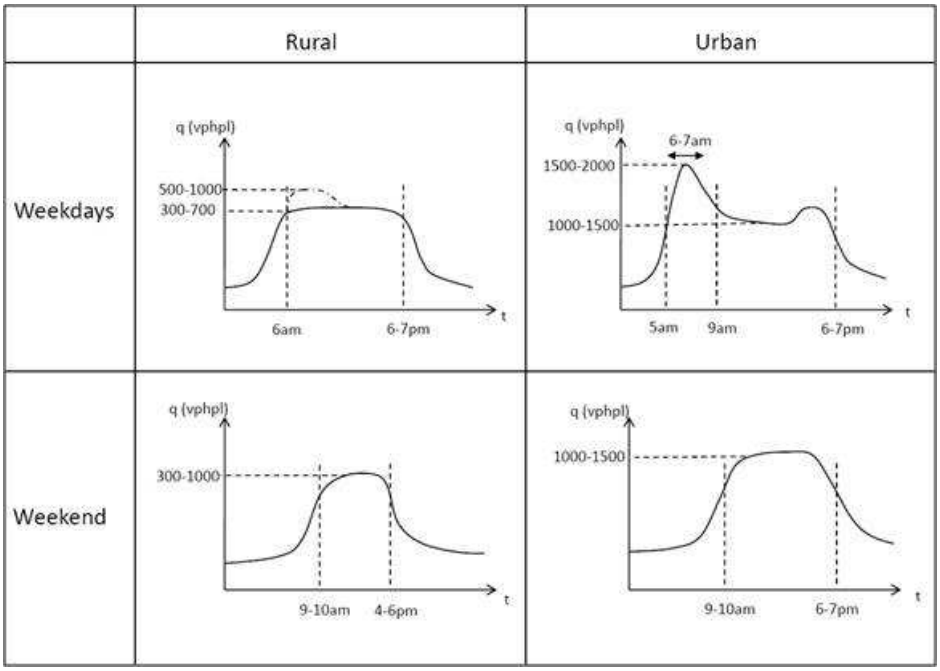


Figure 2: Traffic patterns on urban and rural corridors during weekdays and weekends

Demand generation

This document intents to explain how the OD table is generated for the linear network generated in Paramics. The OD table is a $N \times N$ matrix, where N is the number of zones. Each element T_{ij} of this matrix is the number of trips originated in i with destination j. Different matrices can be defined for different periods of the day.

The number of zones, N, corresponds to the number of entry points (I) plus the number of exit points (J). Note that the entry points include the upstream mainline and all the on-ramps, and the exit points include the downstream mainline and all the off-ramps. That is, $N = I + J$.

Let us number the N zones increasingly in the direction of the flow. Since only one direction of traffic is being analyzed, $T_{ij} = 0$ for $i \leqslant j$. Moreover, $T_{ij} = 0 \ \forall j$ if i corresponds to an off-ramp, and $T_{ij} = 0 \ \forall i$ if j corresponds to an on-ramp ($i = 1..N$ and $j = 1...J$).

We will assume that there are no trips between intermediate ramps. That is:

- all trips generated by the on-ramps are attracted by zone N (i.e. to the mainline downstream end), and
- only trips generated by the first zone (i.e. from mainline upstream) are going to the off-ramps.

In terms of the OD matrix structure, this means that only the first row and last column will have non-zero elements.

Let us define the following variables, which will be defined by the user:

- Q = total number of trips during the simulation period.
- α_k = proportion of the total inflow generated at entry point k, where $k = 1...I$ and $\sum_{k=1}^{I} \alpha_k = 1$ ($k = 1$ corresponds to the upstream mainline). $\bar{\alpha} = [\alpha_1 \ \ \alpha_2 \dots \alpha_I]^T$
- β_m = proportion of the total outflow attracted by exit point m, where $m = 1...J$ and $\sum_{m=1}^{J} \beta_m = 1$ ($m = J$ corresponds to the downstream mainline). $\bar{\beta} = [\beta_1 \ \ \beta_2 \dots \beta_J]^T$

The variables defined fully characterize the demand pattern to be observed in the network. The trips generated at entry point k is given by $\alpha_k \cdot Q$. Likewise, trips attracted

by exit point m is given by $\beta_m \cdot Q$. Note that, given our assumptions, $\alpha_1 = \beta_J$ (the proportion of the total inflow generated by zone 1 is the same as the proportion of the total outflow attracted by the last zone N). Therefore, the user is supposed to provide the scalar value Q and vectors $\bar{\alpha}$ and $\bar{\beta}$.

The variables Q, α's and β's can vary throughout the day. This will define different OD tables and achieve different traffic patterns. By changing the vectors $\bar{\alpha}$ and $\bar{\beta}$ different traffic loads can be given to different ramps, creating merge and/or diverge bottlenecks in the network.

Example: Let us consider a network with $N = 7$ zones, where zones 1, 2, 4 and 5 correspond to entry points ($I = 4$), while 3, 6 and 7 are exit points ($J = 3$). Zone 1 is the mainline upstream end, and zone 7 is the downstream end (see Figure 1).

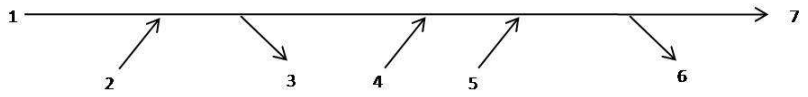


Figure 1: Illustration of the network defined in the example.

Given specific values of Q, α's and β's, the OD table would look as follows:

$$T = \begin{pmatrix} 0 & 0 & \beta_1 \cdot Q & 0 & 0 & \beta_2 \cdot Q & \alpha_1 \cdot Q = \beta_3 \cdot Q \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_2 \cdot Q \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_3 \cdot Q \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_4 \cdot Q \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

If $Q = 10,000$ trips for a 1 hour simulation, $\bar{\alpha} = [0.7 \ 0.05 \ 0.10 \ 0.15]$ and $\bar{\beta} = [0.1 \ 0.2 \ 0.7]$:

$$T = \begin{pmatrix} 0 & 0 & 1,000 & 0 & 0 & 2,000 & 7,000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 500 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1,000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1,500 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Error Model

Purpose

The purpose of this model is to distinguish non-intrusive sensors (radar, acoustic, and video) from intrusive sensors (loops). Non-intrusive sensors are overall cheaper and easier to install, though usually at a cost in terms of data accuracy compared to intrusive sensors.

Methodology

To model the errors made by non-intrusive sensors, we used former studies dealing with sensors errors (appendix). The results given by those studies were extremely disparate. The studies did not all report the same kind of data, for the same conditions, during the same time length, etc. Considering the heterogeneity of the results, we decided to focus on the most precise study [1] and adjusted the model for it to fit all the other studies.

Critical goal

Our objective was to find a model to represent errors made by non-intrusive sensors while calculating speed, count or occupancy. To do so, we determined two traffic conditions: free-flow and congestion and we wanted to have only one model for all non-intrusive sensors.

Model characteristics

- **Free-flow / Congestion**

It was essential for our model to distinguish two conditions, free-flow and congestion. As a matter of fact, almost all studies agree on the fact that sensor results are less accurate during congestion conditions than during free-flow. We chose a single velocity criterion (45 mph) as the limit between congestion and free-flow.

- **Normal distributions**

To model the error, we decided to use normal distributions, which are classic error distributions. The speed error model will use absolute error (in miles-per-hour) and count and occupancy error models will use relative error (in %). Such an approach seems natural and was used in most of the reports ([1], [4], [7]).

- **Mean and Standard deviation**

Independent of the sensor type and traffic conditions, experimental means of errors are rarely zero. However, we did not see a clear tendency in those results, so we thought it was most relevant to take zero as the mean for every model. Consequently, we only needed to determine the appropriate standard deviations for our distributions.

❖ Count and Occupancy

As we could not find any useful quantitative data for occupancy, we compared occupancy and count relative errors for wireless sensors using study [2] (*table 3*), and found that they were similar. Therefore, we chose the same standard deviations for count and occupancy. We also decided that the relative errors for count and occupancy would be correlated. On the other hand, speed errors are derived from an independent random draw.

❖ Interval time length

We considered 5 minute intervals to estimate the error, as it was used in many studies. As it is quite easy to convert standard deviations for different time length, our model could be used for time length of 30 seconds, 1 minute, 2 minutes, 5 minutes, 10 minutes … However, this error model shouldn't be used for time length smaller than 30 seconds. In fact, macroscopic variables like flow, occupancy and speed only make sense at a macroscopic scale. Therefore when the time interval is too small, the meanings of the variables and of the error associated with those variables are lost.

Here is a chart to convert standard deviations for different time length:

Interval = n x 30 sec

	Absolute Error	Relative Error
Speed	x 1/√n	x 1/√n
Occupancy	x 1/√n	x 1/√n
Count	x √n	x 1/√n

❖ Choice of the parameters

To define the standard deviation parameters, we took the results of the most detailed study [1] and adjusted the standard deviation by taking into account that the experimental mean was not zero (*table 2*). Thus, by comparing standard deviations for different sensors and using the order of magnitude of other studies, we chose the standard deviations for our models (*table 1*).

We are aware that better parameters (e.g. for one specific type of sensors) may be used, so we created a configuration file where the standard deviations can be modified.

❖ Lower speeds

For very low speeds (in congestion conditions), our model gives a certain amount of negative speed values. For instance, if the speed is 10 mph, the model gives negative speeds with a probability of 2%. Another way to see this is to notice that negative values are found with a probability higher than 5% for speeds less than 8 mph. In those cases, our model suggests that the computer recalculate the error until a positive speed is found.

❖ Code's functionality

We coded this model (using MATLAB) in order to mimic non-intrusive sensors and add errors to actual speed, count or occupancy. It estimates the error with a random process which follows normal distribution probabilities.

To accurately imitate the sensors, the count is rounded and count and occupancy are estimated based on the same random error. The program iterates the process until positive values are found.

Model

Here is a summary if our results and graphs of the normal distributions representing the error.

***Table 1*: Standard deviations chosen for our model**

5 min interval

Standard Deviations

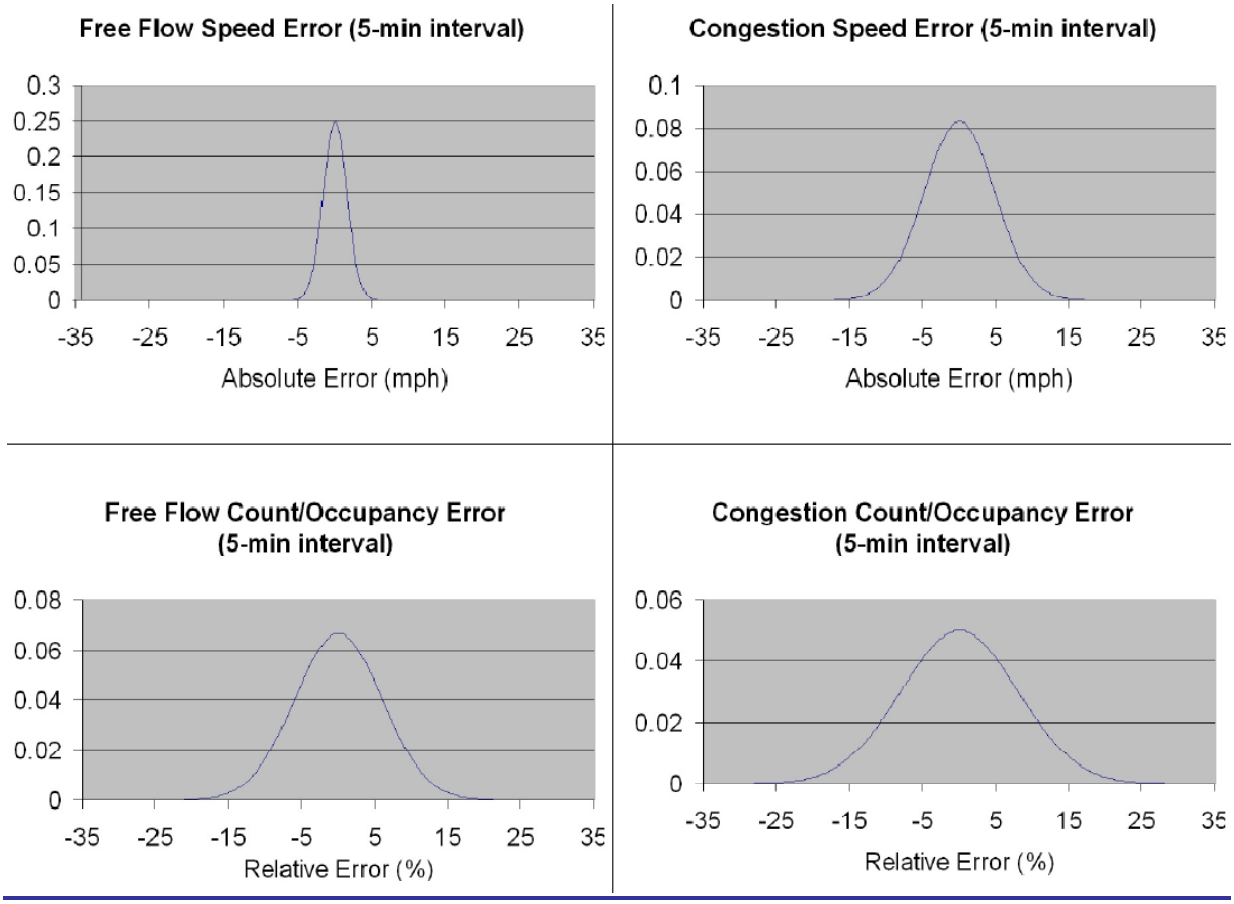
	Abs Error	Rel Error	Abs Error	Rel Error
	free flow		congestion	
speed (mph)	1.6	5.3	4.8	7.4
count (number of vehicles)	8.6	6	11.5	8
occupancy (%)	0.8	6	1	8

30 sec interval

Standard Deviations

	Abs Error	Rel Error	Abs Error	Rel Error
	free flow		congestion	
speed (mph)	5.1	16.9	15.2	23.4
count (number of vehicles)	2.7	19	3.6	25.3
occupancy (%)	2.4	19	3.2	25.3

***Figure 1**: Normal distributions representing the errors for speed, count and occupancy in free-flow and in congestion*



Code

function [speed,count,occupancy] = SensorError(s,c,o,interval)

Given an actual speed s (in mph), occupancy o (percentage between 0 and 100), count c, and a time interval (in seconds) over which everything is measured, this function gives random values that a real non-intrusive sensor might give for speed, occupancy, and count. Since the random values come from a normal distribution, negative values might result. To correct for this problem, if a random value is negative we remove it and obtain a new random value until a positive one is found.

the hard coded values for standard deviations are based on a 30 second time interval

if the time interval is less than 30 seconds then the error model is unreliable; such small intervals must be used for testing purposes only in which case we just return the nominal values given for speed, count and occupancy

```
if interval < 30.0
   speed = s;
   count = c;
   occupancy = o;
   return;
end

SensorErrorConfig;

if s > 45 if speed is greater than 45 mph use free flow parameters
   sigmas = freeflowsigmas/(sqrt(interval/30.0)); standard deviation for speed error term
   sigmaoc = freeflowsigmaoc/(sqrt(interval/30.0)); standard deviation for count and occupancy percent error term
else if speed is less than 45 mph use congestion parameters
   sigmas = congestionsigmas/(sqrt(interval/30.0));
   sigmaoc = congestionsigmaoc/(sqrt(interval/30.0));
end

error = random('norm',0,sigmas);
while error < -s
   error = random('norm',0,sigmas);
end
speed = s + error;

percenterror = random('norm',0,sigmaoc)/100.0;
while percenterror < -1
```

```
    percenterror = random('norm',0,sigmaoc)/100.0;
end
occupancy = o*(1+percenterror);
count = round(c*(1+percenterror));
if count == 0 && occupancy > 0
    count = 1;
end
```

Configuration file for SensorError function

contains standard deviation values for speed, occupancy, and count parameters for free flow and congestion

```
freeflowsigmas = 5.06;
freeflowsigmaoc = 18.9;
congestionsigmas = 15.2;
congestionsigmaoc = 25.3;
```

Appendix

List of the studies:

- [1]: *Vehicle Detector Evaluation*, Texas Transportation Institute, 2002
- [2]: *Evaluation of Wireless Traffic Sensors by Sensys Networks*, Inc., CCIT, 2006
- [3]: *Traffic Data Collection Methodologies*, Pennsylvania Department of Transportation, 2006
- [4]: *An Assessment of Loop Detector and RTMS Performance*, California PATH Program, 2004
- [5]*Evaluation of Non-intrusive Technologies for Traffic Detection*, Minnesota DOT, 2002
- [6]: *Video detection System Testing*, URS Corporation, 2003
- [7]: *Detector Technology Evaluation*, University of Utah, 2003

***Table 2**: Results of 'Vehicle Detector Evaluation, Texas Transportation Institute, 2002' study*

Speed (miles per hour)

		free flow	congestion
Radar	mean	0.83	6.15
	std dev	1.39	4.58
	RSME	1.62	7.67
Video	mean	0.12	-1.62
	std dev	0.74	1.83
	RSME	0.75	2.44
Acoustic	mean	0.59	-9.44
	std dev	2.22	8.71
	RSME	2.3	12.84

Count (number of vehicles/5-min)

		freeflow	congestion
Radar	mean	-2.45	-3.16
	std dev	3.29	6.03
	RSME	4.1	6.8
Video	mean	-11.04	-8.88
	std dev	10.47	5.02
	RSME	15.21	10.2
Acoustic	mean	-9.27	-9.44
	std dev	4.45	8.71
	RSME	10.28	12.84

***Table 3*: Results of the 'Evaluation of Wireless Traffic Sensors by Sensys Networks, Inc., CCIT, 2006' study.**

Accuracy-Reference to loops for 5-min samples

Measurement	Unit	Average	Average Bias	Average StdDev	Absolute RSME	Relative RSME
Occupancy	%	12.5	0.2	0.5	0.6	4.7
Count	number of vehicles / 5 min	92.5	1.7	3.7	4.1	4.4
Speed	mph	48	0.7	0.8	1	2.1

Accuracy-Reference to loops for 30-sec samples

Measurement	Unit	Average	Average Bias	Average StdDev	Absolute RSME	Relative RSME
Occupancy	%	12.5	0.2	1.4	1.4	11.5
Count	number of vehicles / 30 sec	9.5	0.2	1.1	1.1	11.8
Speed	mph	48	0.7	2.3	2.4	5

Detector requirements for ramp metering in California

Lianyu Chu
California Center for Innovative Transportation (CCIT)

Introduction

Ramp metering is the major traffic control strategy from the freeway traffic management agency. The benefits of ramp metering are:

(1) Restrict the total flow entering the freeway by temporarily storing some traffic on the ramps in order to ensure that mainline freeway is operated within the freeway's capacity and prevent congestion.
(2) Break up platoons of vehicles entering freeways in order for vehicles from onramps to merge more easily and provide safety.
(3) Divert some vehicles to other routes due to the waiting time and thus reduce demand going to the freeway.

Existing detector placement

California has widely applied ramp metering in major metropolitan areas.
There are currently three ramp metering systems applied in California, which are summarized in Table 1. San Diego Ramp Metering System (SDRMS) is widely used in Districts 3, 6, 8, and 11 (1). Semi-Actuated Traffic Management System (SATMS) is used in Districts 7 and 12 (2). Traffic Operations System (TOS) is currently deployed in District 4 (3). All these three ramp metering systems are local traffic responsive control and they need to have a mainline detector placed upstream of the on-ramp. The theory behind SDRMS and TOS is occupancy control and the theory behind SATMS is demand capacity control. .

Table 2 Existing Ramp Metering Systems

Existing systems	Districts	Theory behind	Mainline Detector
SDRMS	3, 6, 8, 11	Occupancy control	Upstream Detector
SATMS	7, 12	Demand capacity	Upstream Detector
TOS	4	Occupancy control	Upstream Detector

Ramp Meter Design Manual from California Department of Transportation (Caltrans) has the following statement. “Caltrans is committed to using ramp metering as an effective traffic management strategy to maintain an efficient freeway system and protect the investment made in constructing freeways by keeping them operating at or near capacity. Ramp Metering is an integral part of the Traffic Operations Program Strategic Plan which outlines the program’s commitment to focus first on implementing operational strategies to reduce congestion and increase safety on California’s state highway system.”

Ramp metering control usually needs to have detectors to be installed to the freeway mainline and ramps. Table 1 summarizes the detector placement requirement based on Caltrans Ramp Meter Design Manual. Figure 1 shows the requirement in a figure.

Table 2 Detector placement requirement based on Ramp Meter Design Manual

Detector	Placement requirement
Mainline Detector	(1) Two loops per lane should be installed on the mainline. (2) Spacing shall be 6.1 m from leading edge to leading edge (3) Located upstream of the entrance ramp nose, opposite the limit line.
Ramp Detector	(1) Ramp loops (demand and passage) should be installed for each entrance lane near the limit line. (2) The number and spacing of ramp loops should be determined by the District Operations Branch responsible for ramp metering - District 11: typically 4 demand loops - District 3,4,8: typically 3 demand loops - District 7, 12: typically 2 demand loops
Exit Ramp Detector	One loop per exit ramp lane should be installed for count information and loop calibration.
Queue Detector	One loop per entrance ramp lane should be installed for queue detection near the connection of the surface street.

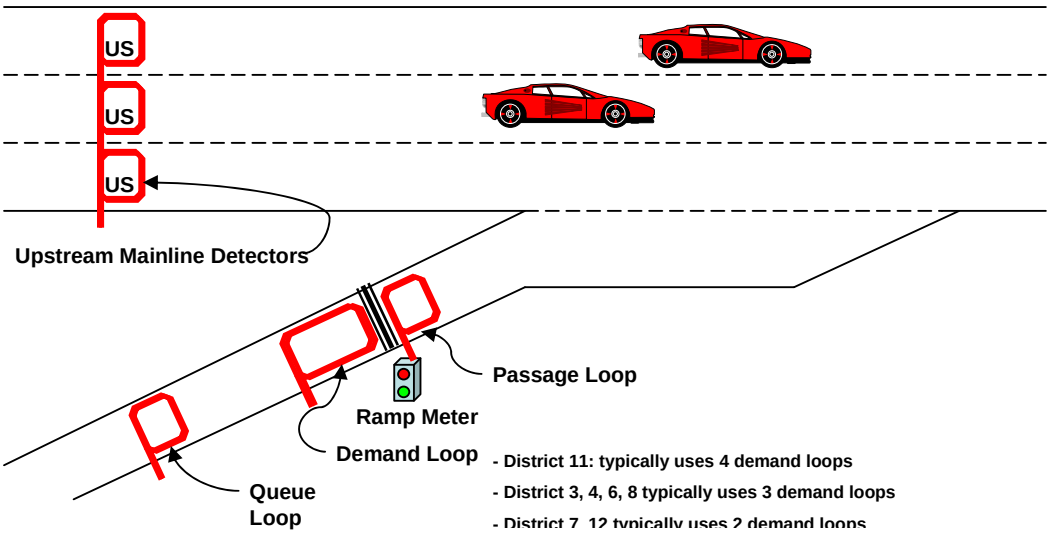


Figure 1 Detector placement around a meter

According to the real detector placement practice, Caltrans usually install one more detector to the downstream of the passage detector. The detector is called on-ramp detector. For those areas (such as District 7 and 12) with HOV bypass on-ramp lanes, there is a HOV detector to be placed on the bypass lane. Both detectors are placed for count information and loop calibration. Figure 2, which is from District 7's ATMS Traffic Engineer's Manual, shows the typical detector placement around a ramp in Districts 7 and 12.

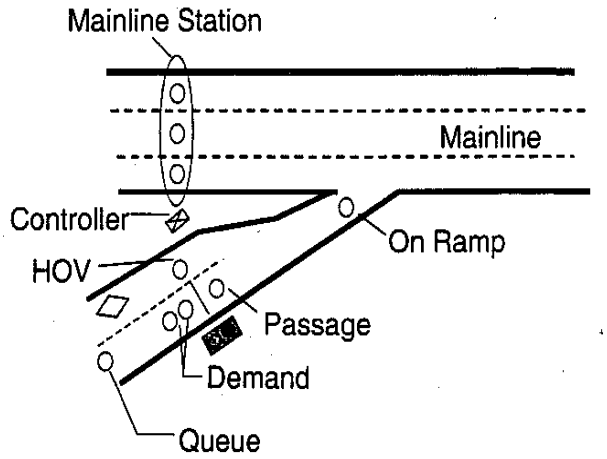


Figure 2 Typical detector placement with HOV bypass lane in District 7 and 12

SWARM Ramp metering

Since late 1990s, Caltrans has started testing the System-wide Adaptive Ramp Metering (SWARM), developed by NET or Delcan (4,5,6). SWARM is a central ramp metering algorithm embedded in Caltrans' TMC software Advanced Transportation Management System (ATMS). SWARM has four algorithms:

(1) SWARM 2a: Point detector based algorithm
(2) SWARM 2b: Segment wide algorithm
(3) SWARM 2c: Mainline detector based algorithm
(4) SWARM 1: System wide algorithm

SWARM is a centralized metering system at TMC. For ramp meters in the field, they are directly controlled and operated by local traffic controllers running the SATMS ramp metering program. SWARM can not be setup without SATMS. SATMS needs mainline upstream detector. The minimum detector requirement for SWARM is that each meter must have a corresponding mainline upstream detector.

SWARM 2a

SWARM 2a uses only one detector, which is defined as located upstream of the onramp on the mainline.

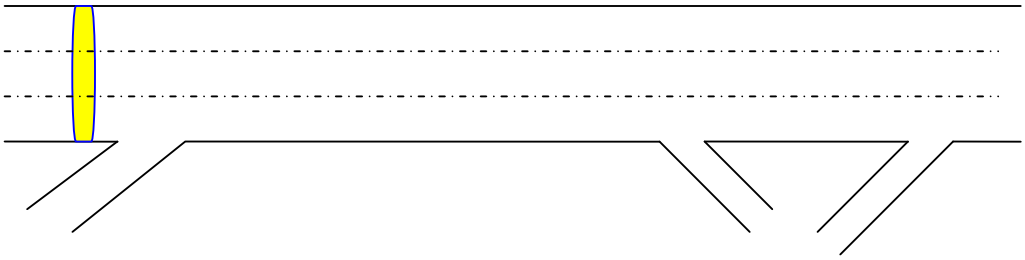


Figure 3 SWARM 2a detector requirement

SWARM 2b

SWARM 2b needs to define a target section on the freeway mainline. The target section starts from the current onramp's corresponding mainline detector and ends at the next available mainline detector with good data. If there are any onramps and offramps within the two detectors, there must be detectors on onramps and offramps. For a typical section that has both onramp and offramp, the ideal downstream detector can be placed immediate downstream of the offramp.

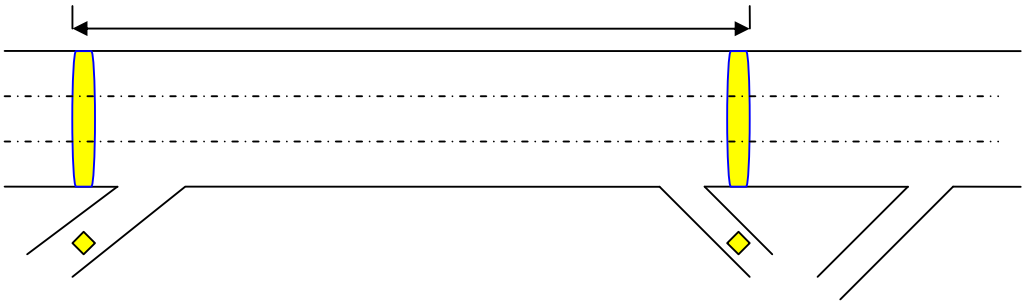


Figure 4 Definition of the target section

The target section may have one aux lane. In order for SWARM 2b to work with this situation, SWARM 2b should have a local parameter to consider the effects of the aux lane.

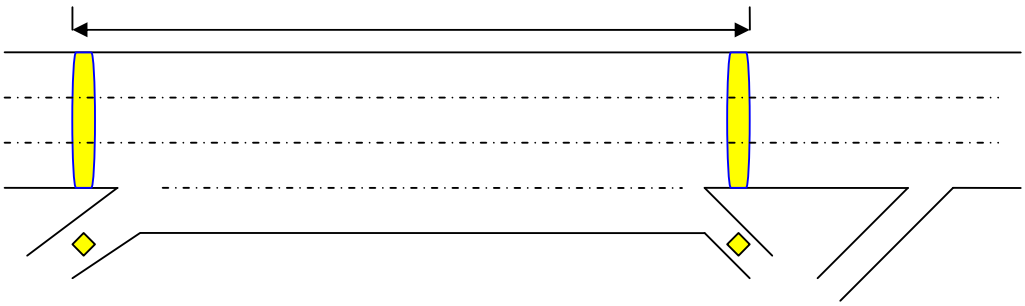


Figure 5 Definition of the target section with an aux lane case

The location of the downstream mainline detector can be further relaxed. It can be the mainline detector of the next onramp. This is a normal detector configuration in California's urban freeways.

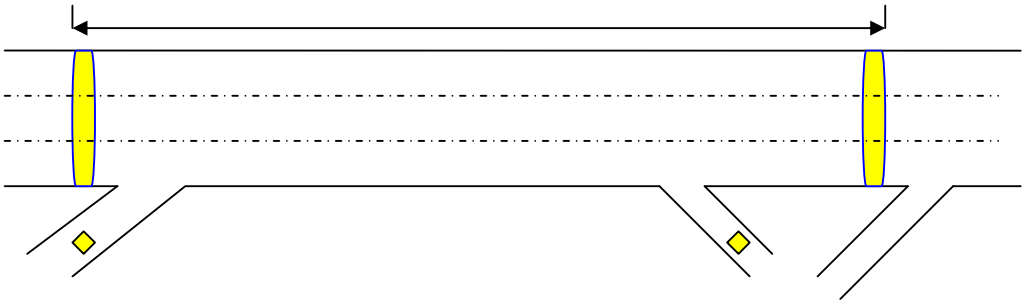


Figure 6 Target section based on the normal detector configuration

In order for SWARM 2b to perform, the basic assumption is that the traffic condition in the target section is homogenous or the traffic density for each subsection of the target section is similar. This is a tough assumption that won't be satisfied if there is

any queue within the section. As a result, it will be beneficial if there is an additional detector between the onramp and offramp. As a result, the target section becomes smaller.

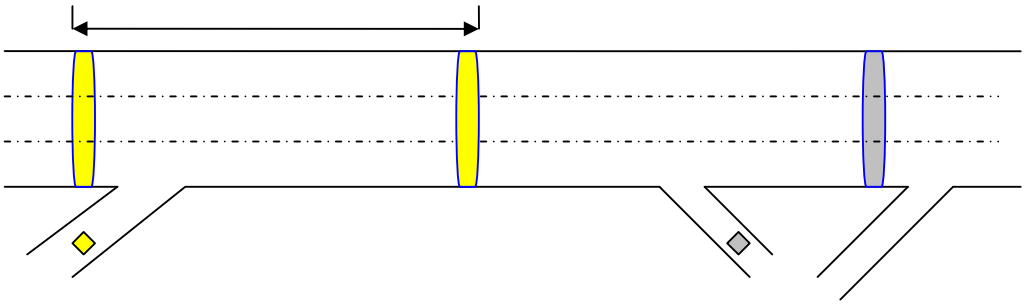


Figure 7 One more mainline detector between onramp and offramp

There is a tradeoff on the spacing of the two mainline detectors. If they are close, SWARM 2b can only use the traffic data of a small portion of the mainline to determine its traffic condition and metering rate. If they are far away, the section is bigger and it may not be a homogenous section and thus metering rate from SWARM 2b is not accurate. So, the two mainline detectors should have a certain distance in order to provide a good area wide measure of traffic condition of the target section.

The direct benefit of ramp metering is to delay the occurrence of traffic congestion. From this perspective, the mainline detector should be placed at a location where the shockwave is usually initiated. As we know, some drivers change lane when they see the guidance sign. So, the detector may need to place at the location that drivers can start seeing the guidance sign.

If there is more than one mainline detector within the definition of the target section as shown in Figure 6, it won't help but will make the actual target section to be small. It may negatively affect metering performance.

If there is any lane drop (which doesn't mean the end of onramp acceleration lane) within the definition of the target section as shown in Figure 6, a mainline detector is strongly suggested being placed downstream of the lane drop point. Please note this is my thought and where to place mainline detector for the lane drop case may need to be further discussed together with Henry and you guys based on simulation results.

.

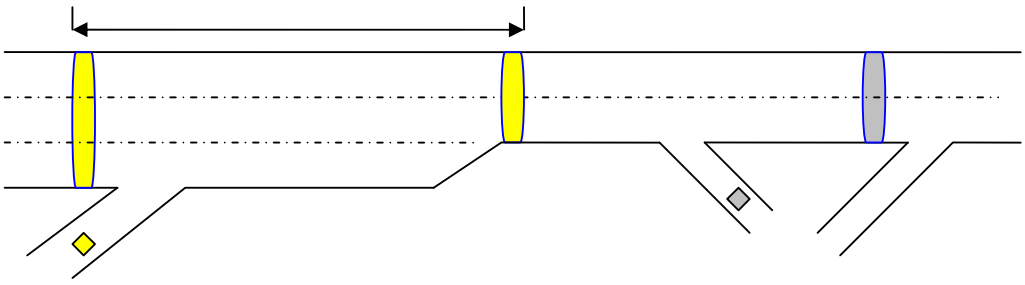


Figure 8 Lane drop case

Similarly, if there is any lane gain (slip lane for off-ramp doesn't count) within the definition of the target section as shown in Figure 6, a mainline detector can be placed

upstream of the lane gain point in order to enhance the accuracy of the SWARM 2b. But, this requirement is not as critical as the lane drop case.

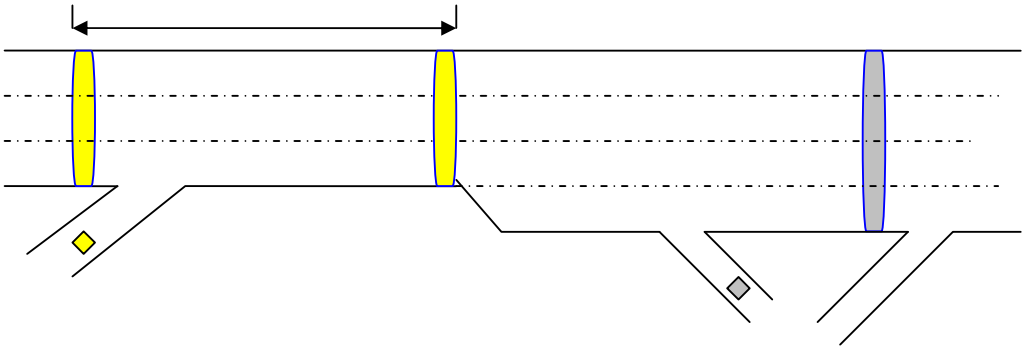


Figure 9 Lane gain case

SWARM 2c

The theory behind SWARM 2c is coordinated occupancy control. SWARM 2c needs to average and smooth the occupancy data of mainline detectors upstream and downstream of the on-ramp, and then lookups the metering rate from a pre-determined occupancy-metering rate table, pre-determined based on an traffic flow analysis of historical traffic data.

SWARM 2c further relaxes the detector requirement. It only needs two mainline detectors. For the normal detector setting for urban freeways in California, the first one is the current onramp's corresponding mainline detector and the second could be the immediate downstream mainline detector with good data.

SWARM 2c has similar requirement for the definition of the target section. Compare to SWARM 2b, SWARM 2c removes the requirement for onramp and offramp detectors. Except the onramp and offramp detector requirements, the discussion about SWARM 2b's detector requirements also apply to SWARM 2c (i.e. Figure 7-9).

SWARMS 2c evaluates the traffic condition of the target section based on the occupancy values of both mainline detectors. The occupancy from the upstream detector and the one downstream are averaged and smoothed and then used to lookup the metering rate from a local table.

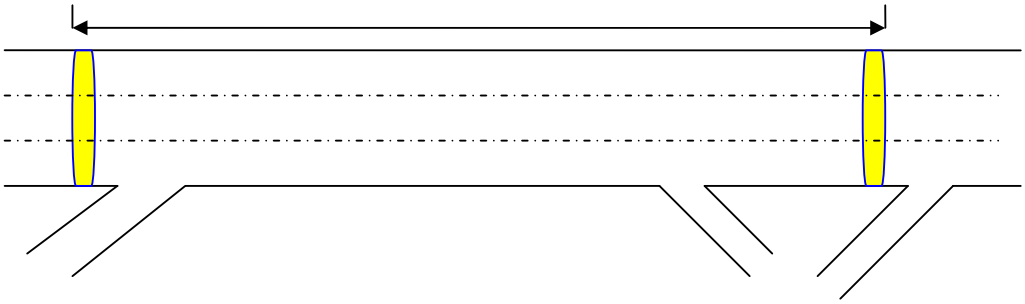


Figure 10 Detector requirements for SWARM 2c

SWARM 1

SWARM 1 needs to have mainline detectors to be placed at all bottleneck locations and counting detectors on all onramps.

As shown in Figure 11, there are two bottlenecks for the network. Those detectors in yellow color are required by SWARM 1.

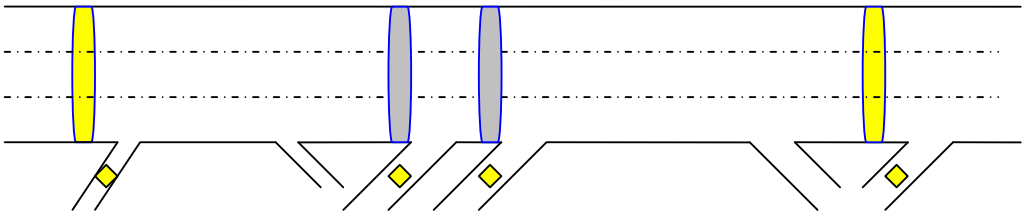


Figure 11 Detector requirements for SWARM 1

In order to determine bottleneck locations, University of Minnesota's work can be directly used.

Summary

Table 3 summarizes the basic parameter and the mainline detector requirements for each SWARM algorithm.

SWARM needs a mainline detector to be placed upstream of each onramp. In order for SWARM 2b and 2c to work appropriately, as shown in Figure 3a, SWARM requires a detector located upstream of the ramp and another located downstream of the ramp. As shown in Figure 3b, the mainline detector upstream of the next on-ramp can be used as the "downstream" detector. However, this relaxation will not apply if thee is a bottleneck, caused by either lane drop or strong weaving or merging, between the two mainline upstream detectors. The placement of a detector at bottleneck location is also expected by SWARM 1.

Basic parameters of SWARM are either density or occupancy. The calculation of density depends on occupancy. Hence the sensors should be deployed to provide the "best" estimate of the mainline occupancy.

Table 3 SWARM's mainline detector requirement

	Algorithm parameter	ML Detector requirement
SWARM 1	Density	Detector at bottlenecks
SWARM 2a	Density	Corresponding upstream detector
SWARM 2b	Density	(1) Corresponding upstream detector (2) A downstream detector, which may be the next on-ramp's upstream detector, or a detector in between.
SWARM	Occupancy	(1) Upstream detector

2c		(2) A downstream detector, which may be the next on-ramp's upstream detector, or a detector in between.

References

(1) Caltrans Traffic Operations Program, Ramp Meter Design Manual, January, 2000.
(2) Caltrans, SATMS 170 controller Program Users Instruction, Nov 25, 1986.
(3) Caltrans District 4 (2004) TOS v2.1.1 User Instructions, 2004.
(4) System Wide Adaptive Ramp Metering - High Level Design, Final Draft, Prepared by NET for Caltrans and FHWA, June 19, 1996.
(5) Advanced Transportation Management System Traffic Engineer's Manual, Revision 1, Prepared by NET for Caltrans District 7, June 2000.
(6) Advanced Traffic Management System (ATMS) Release 2 - Code Specifications (540) - SWARM, prepared by NET, 4/9/2003.
(7) H. Pham, et al (2002) SWARM Study Final Report on W/B Foothill Freeway (W/B LA 210), Research Report to California Department of Transportation.
(8) Michael Zhang, Taewan Kim, Xiaojian Nie, Wenlong Jin, Lianyu Chu, and Will Recker. Evaluation of On-ramp Control Algorithms, Research Report: UCB-ITS-PRR-2001-36. California Partners for Advanced Transit and Highways. 2001.

SECTION 3 – ANALYTICAL FRAMEWORK

X. BAN, R. HERRING, JD MARGULICI, J.C. HERRERA, A. BAYEN. “DEVELOPMENT OF A SIMULATION FRAMEWORK FOR INVESTIGATING OPTIMAL DETECTOR SPACINGS FOR FREEWAY TRAVEL TIME ESTIMATION”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2007.

Development of A Simulation Framework for Investigating Optimal Detector Spacings for Freeway Travel Time Estimation

1 Introduction

Travel time estimates on selected itineraries arguably constitute the single most relevant roadway traffic metric. First, travel time is a crucial measure of traffic conditions and system performance. Travel time reliability has been receiving particular attention from FHWA lately, and recent research aims to quantify trip time reliability in a network level [1], [2], [3]. Second, travel times represent information that is easy for the driving public to understand and process. Numerous studies [4], [5], [6], [7], [8] reveal that commuters appreciate and value travel time information, which reduces their uncertainty and their stress. Further, relevant information can arguably enable travelers to make educated choices about their itinerary, departure time or even transportation mode, with the result of bringing about a form of "system self-management."

Enjoying a poster child status in the ITS industry, travel time estimates have benefited from a flurry of innovations in traffic data collection, processing techniques, and information delivery modes over the past decade. On the front end, both government agencies and private media ventures across the world's largest cities provide traffic information and travel time estimates through a variety of channels, including web browsing, traditional and satellite radio, mobile devices, navigation units and, increasingly, electronic signage on roadways.

Despite those technical improvements, accurate and timely travel time estimates remain a rare commodity. Even in the San Francisco Bay Area, where one of the best 511 traveler information systems in the United States has been deployed, anecdotal evidence suggests that real-time travel time estimates are quite often off-the-mark. Note that anecdotal evidence is invoked in place of

systematic studies because those are surprisingly sparse [9]. One reason for that may be that the industry has not developed widely accepted metrics and methods to measure the quality of travel time estimates. Most benchmarks involve a few "probe runs" that are assumed to represent "ground truth", leaving serious doubts regarding the statistical validity of the results. Based on the emerging availability of probe-based data collection techniques such as license plate readers, electronic toll collection tag readers and cell phones or Personal Digital Assistants tracking, this report suggests metrics to gauge the quality of travel time estimates. More precisely, we look into two types of travel time estimates: real-time predictive estimates (such as *instantaneous* travel time estimates described in this report), which attempt to predict the median travel time on a segment at the time of entry, and travel time reconstruction estimates (such as *dynamic* travel time estimates described in this report), which use passed data and are useful as performance measures.

Measuring the quality of travel time estimates is important for the following reasons: 1) the margins of errors of travel time estimates should be better understood and formulated so that drivers and operators can develop adequate expectations; 2) robust validation and monitoring practices for travel time estimates can point to needed improvements in traffic data collection and ultimately build up the confidence of network operators in the information; 3) in the context of public-private partnerships for data collection, aggregation and dissemination, quality metrics are needed to enable government agencies and technology providers to reach business agreements.

When one estimates real-time predictive travel times, there is obviously a part of unpredictability that can never be eliminated. Even the best traffic model is no crystal ball for future incidents. But if we abstract from that inherent constraint, travel time estimates quality depends on the quality of the data that is collected to estimate it, and to some extent on the strength of the underlying model that is used. Obeying the "garbage-in, garbage out" adage tells us that data quality is a prerequisite, whereas the model is a refinement. The premise of this study is that travel times can be estimated with fairly good accuracy if the data is abundant enough. In other words, we view the problem of estimation as one of information. Estimation errors are driven by the information gap between real traffic conditions and what can be inferred from detector data. Clearly, if operators had access to the exact parameters of every single car on a freeway in real time, then the estimation of predictive travel times would only be a matter of selecting the best traffic model, and reconstructed travel times would be known and not require estimation. In reality, detectors only provide a sample of information. The coverage and reliability of this sample send travel time estimates quality up or down.

In order to measure the information gap between ground truth and detector data, this study

considers various traffic flows for which vehicle trajectories are known. Some of those traffic flows are constructed theoretically from the fundamental diagram and the car following model assumption. They include an oscillatory congested flow and a bottleneck model. Those models do not pretend to be realistic representations of traffic but they provide an analytical framework in which the discrepancies between ground truth and detector data can be tracked and understood. Additionally, the study considers real trajectories collected by video and machine vision techniques at the Berkeley Highway Lab (BHL) as part of the NGSIM (Next Generation SIMulation) project.

In the past, many studies have focused on algorithmic techniques to estimate travel time from available field data, which generally came from inductive loops [10], [11]. Therefore, most studies assumed detector locations to be given and proposed ways to optimally use the data. This study takes a different stand. Minimal attention is given to algorithmic techniques, though we acknowledge they may be useful. Instead, this report assumes relatively simplistic algorithms (which in effect are the ones used by transportation operators worldwide) and focuses on the relationship between detector deployment (such as locations and spacing) and travel time estimation quality.

There are two operational questions that guide this study. The first question is one that many practitioners today would like an answer to: given the number and locations of traffic detectors on a corridor, how reliably can travel times be estimated? Typically, an operator that has installed a Changeable Message Sign on a freeway corridor needs to know with a high degree of confidence that detection downstream of the sign is adequate before deciding to display travel times on it. By studying the sensitivity of travel time estimates to detector density, this report provides a first-order answer to that question. The second question concerns the fortunate but maybe puzzled operator who is given the budget to add detectors on a corridor. In that case, the operator needs to determine the best detector deployment configurations (i.e., detector types, spacing and locations) that can result in proper travel time estimation in a cost-effective manner. Again, this report provides guidance by exploring the relationship between detector configuration and travel time estimation quality, a topic that has been studied only recently [12], [13].

Two additional efforts will complement this report to fully assess the relationship between detector configuration and travel time estimates quality. The first effort will continue the present analysis but be based on much more sophisticated flows generated through micro-simulation. Such flow models will carry more realism than the theoretically-generated flows used in this report. The second effort will be an empirical study, implying a completely different approach in which trajectories are not known, but at least some travel times can be observed and compared to estimates obtained from various sets of traffic detectors.

2 Methodology Overview

As aforementioned, the purpose of our study is to explore the trade-off between detector deployment and the quality of travel time estimation. First, detectors may only collect imperfect traffic information. This is because 1) Information from individual detection station is imperfect because detectors are error-prone, 2) The data transmitted by the station is almost always aggregated, and 3) The density of stations will critically impact the overall quality of traffic information for a given corridor. For example, loop detectors usually record the average aggregated traffic speed at a given location for a fixed time period (usually 30 seconds). One may thus expect that travel times estimated using data from detectors very likely deviate from the "true" travel times. How large the deviation is will heavily depend on detector deployment configurations. With more detectors (i.e., smaller detector spacing), the estimation quality may increase, but the deployment cost will increase as well. Therefore, it is crucial to determine the most cost-effective detector deployment strategies that can result in appropriate travel time estimation performance.

Determining the optimal detector deployment, including at least detector locations and spacing, is not trivial. In this report, we take the first step towards this task, i.e., to explore the relationship between detector deployment and travel time estimation quality. In particular, we look at, for a given segment of freeway, how travel time estimation quality depends on detector spacing. To be more tractable, we only investigate deployment scenarios that detectors are evenly spaced. We further vary detector spacing from very sparse (such as one detector every 2-3 miles) to very dense (such as one detector every 0.1 mile) and study the trend of travel time estimation quality as detector spacing changes. It is the authors' understanding that this will hopefully give us the first-hand knowledge regarding travel time estimation and detector spacing, which will be an important input to more rigorous studies on optimal detector deployment strategies.

To assess the quality of travel time methods, we need to know both the "ground-truth" travel times and travel time estimates. The former can be obtained via perfect knowledge about vehicle trajectories. This may sound a strong assumption in the first glance, however, there are (at least) three ways one can obtain trajectory data. First, some advanced traffic detectors, such as video cameras, can produce vehicle trajectory data directly. For example, NGSIM [14] provides trajectory data for individual vehicles for several locations in California. However, trajectories obtained this way still requires a very high experimental cost and on a limited segment length nowadays. Second, one can "reconstruct" vehicle trajectories from other traffic measurement (like speeds) by applying certain traffic flow and car following theory, as will be discussed later in this report. The third way to obtain vehicle trajectory data is through micro-simulations. Since micro-simulation tracks the

detailed movement of individual vehicles, vehicle trajectories can be readily generated via running micro-simulation. Given vehicle trajectories, ground-truth travel times can be easily obtained, as discussed in Section 2.1.1.

The estimated travel times will depend on which travel time methods to use. In this report, we focus on two types of travel time estimates: real-time predictive estimates (such as *instantaneous* travel time estimates), which attempt to predict the median travel time on a segment at the time of entry, and travel time reconstruction estimates (such as *dynamic* travel time estimates), which use passed data and are useful as performance measures. Both methods assume that aggregated speeds (e.g., over 30-second periods) are the only data that are available from detectors, which are again computed using vehicle trajectories. These two methods are selected because they are the simplest and most widely used travel time methods by practitioners. Section 2.1.2 provides detailed descriptions on how to compute the instantaneous and dynamic travel times based on vehicle trajectories.

Another critical component in evaluating travel time estimation is quality measures. In this report, we define two metrics to capture the accuracy and relevance of travel time estimates with respect to the ground-truth travel times. Both measures are based on the relative errors of estimated vs. actual travel times of individual drivers. They are, however, aggregated measures of multiple drivers in a pre-defined time period. The accuracy measure focuses on the average trend of estimated travel times, represented by the mean relative error of travel time estimates. The relevance measure captures the variation of travel time estimates, defined as the 75-th percentile absolute value of the relative errors of travel time estimates over ground-truth travel times. Discussions of travel time quality measures are presented in Section 2.2.

In this report, we concentrate on reconstructing vehicle trajectories from specific assumptions regarding traffic states (i.e., the second method to obtain trajectories). This is particularly done in Sections 3 and 4 for wave propagation and oscillatory traffic states, respectively. These two flow models are selected because they are probably the simplest traffic flow states after the free flow condition. Wave propagation state corresponds to places upstream of active bottlenecks where congestion starts to form and propagates upwards. Oscillatory traffic, on the other hand, represents stop-and-go traffic in heavy congestion. Therefore, the two traffic flow states have significant applications in practice. In this report, we will simplify these two flow states to the extent possible. The purpose is to capture their most basic features so that the nature of the relationship between detector deployment and travel time estimation quality can be revealed.

The way we conduct the investigations in this report is mainly through numerical studies. That

is, we select typical parameters for each of the studied traffic states. We then deploy imaginary vehicles for a period of time, using headway either fixed as a priori or determined using the fundamental diagram. Based on certain traffic flow models, we can obtain vehicle trajectories for a set of vehicles, which will also provide the vehicle actual travel times. For each vehicle, the estimated travel time can be calculated given a detector deployment scenario. The accuracy and relevance measures can then be computed using the estimated and actual travel times for all vehicles within the given time period. The quality measures will be calculated for different detector spacing scenarios (all evenly spaced), through which the relationship between travel time estimation quality and detector spacing can be achieved.

2.1 Travel Times from Vehicle Trajectories

2.1.1 Actual Travel Times

Given vehicle trajectories, the travel time of every vehicle on a given section is directly accessible. As depicted in Figure 1, the solid lines represent trajectories of individual vehicles in the "space-time" diagram (i.e., the $x-t$ diagram). Suppose that we are interested in the travel time of route r between two points A and B, as shown by the bold line along axis x. Then, for a vehicle entering the route at time t and leaving at time t_{out}, the actual travel time of the vehicle at t, denoted as $\tau_{\text{act}}(t)$, will be:

$$\tau_{\text{act}}(t) = t_{\text{out}} - t. \tag{1}$$

2.1.2 Estimated Travel Times

There are numerous ways available in the literature to compute estimated travel times. In this research, we focus on two simple and widely used methods: real-time predictive estimates and travel time reconstruction estimates. The former is also called *instantaneous travel time* in this report, which assumes traffic conditions remain unchanged from the time a vehicle enters a route until it leaves the route. Therefore, the travel time of the route can be computed by summing the travel times of the constituent links at the time a vehicle enters the route. The latter, also called the *dynamic travel time* in this report, is also a summation of travel times of its constituent links; however, each link travel time will be computed using the latest traffic condition at the time a vehicle enters the particular link. Therefore, arguably, the dynamic travel times match reality

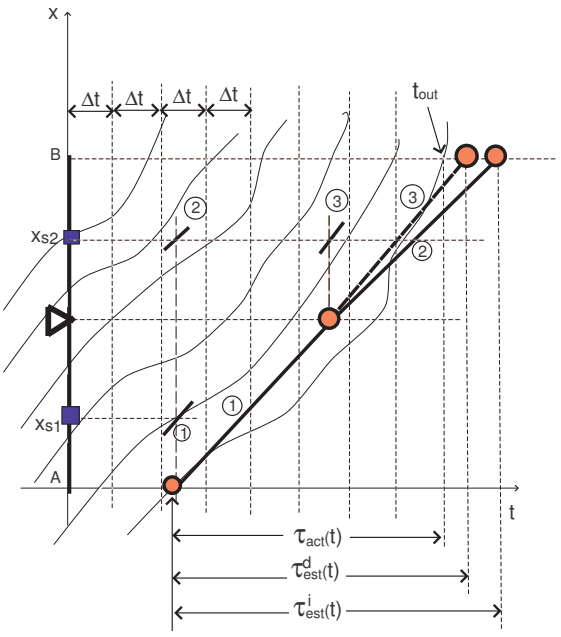


Figure 1: Travel Times from Vehicle Trajectory

more closely because it uses the most recent available information. So it should result a better estimate of the ground-truth travel time.

Figure 1 illustrates how instantaneous and dynamic travel times are effectively computed from vehicle trajectories. Suppose that there are two speed detectors (marked as small squares on route r), located at x_{s1} and x_{s2}. We assume these two detectors divide the route r into two links, separated by the solid triangle symbol in the middle of the two detectors. A link defined in the this manner will be associated to only one detector[1]. Note that this definition is widely used in practice (see for example [15]). We further assume that at any given time instant, a link will experience the speed that is collected by its associated detector.

In practice, a detector generates the average vehicle speed over a pre-defined time period, Δt (Δt=20 or 30 seconds in many applications). Hence, the average speed recorded by a detector at time t can be obtained from trajectories of individual vehicles passing the detector during the time period Δt that encompasses t. Assume there are m of such vehicles and v_i is the speed of the i-th vehicle passing by the detector. Mathematically, the average speed of detector s at time t, denoted as $v_s(t)$, can be computed as:

[1]There are other ways to define links, e.g., a link can be defined as the segment between two neighboring detectors. It is our understanding that all these definitions should be equivalent since they are just (slightly) different ways to use the same information (i.e., detector data)

$$v_s(t) = \frac{\sum_{i=1}^{m} v_i}{m}. \tag{2}$$

In equation (2), the average detector speed at time t is actually the time-mean speed of all vehicles passing the detector during the sampling time period that encloses t.

With detector speeds in place, the travel time of the link associated with a detector can be computed as:

$$\tau_a(t) = L_a / v_{s_a}(t), \tag{3}$$

where s_a is the location of the detector associated with link a and L_a is the length of the link. Then the instantaneous travel time of a route can be computed as:

$$\tau^i_{\text{est}}(t) = \sum_{a=1:A} \tau_a(t). \tag{4}$$

Here $\tau^i_{\text{est}}(t)$ denotes the instantaneous route travel time at time t, $\tau_a(t)$ is the instantaneous travel time of the a-th link ($1 \le a \le A$), and A is the number of constituent links of the route. In Figure 1, the bold solid line illustrates the calculation of the instantaneous travel time for route r at time t.

Similarly, the dynamic travel time of a route, denoted as $\tau^d_{\text{est}}(t)$, is:

$$\tau^d_{\text{est}}(t) = \sum_{a=1:m} \tau_a(t_a), \tag{5}$$

where $t_a = t + \sum_{b=1:a-1} \tau_b(t_b)$ is the entrance time of a vehicle to link a.

In Figure 1, the bold dash line represents the calculation of the dynamic travel time of route r at time t. Notice that except for the first link, the associated detector speed and link travel time are updated at the entrance time to any other constituent links of the route.

2.2 Quality Measures Definition

Quality measures are quantitative metrics for evaluating the quality of estimated travel times. Measuring the quality of travel time estimates is complicated by the fact that there is no single

trip time value on a given road segment at a particular time. Not all vehicles travel at the same speed, and drivers experience different trip times as a result. However, for practical purposes, an estimate that captures the likely trip time of most vehicles is provided (some public agencies provide a range instead of a single value: this may be a better way to communicate expectations, but the range is still based on a baseline estimate). Each individual driver observes two values: a travel time estimate, and his or her actual trip time. Therefore, it is possible to compute an individual relative error between those two values, defined as the ratio of their difference to the actual trip time. Individual relative errors should ideally be as close as possible to zero. They deviate because 1) the estimate may be biased and 2) certain individuals travel slower or faster than most other drivers. The second factor is not controllable but will nonetheless affect the perceived usefulness of the estimate to those individuals. As a result, we assemble two metrics: accuracy and relevance. Note that these two measures are mainly for benchmark travel time methods. To study the the relation between detector deployment and travel time estimates and hence the optimal detector locations, a single measure that combines both the accuracy and relevance (reliability) may be more preferable. This will be reported in a separate document.

2.2.1 Accuracy Measure

The *accuracy measure* assesses how close the estimated travel time of a generic motorist is to his/her actual travel time. To define the accuracy measure, we first define the relative error. Assume the i-th driver's actual trip time at is τ_i and the estimated travel time is $\hat{\tau}_i$. Then the relative error, denoted as e_i, is defined as:

$$e_i = \frac{\hat{\tau}_i - \tau_i}{\tau_i}. \tag{6}$$

The accuracy measure is then defined on a pre-defined time period $[T_1, T_2)$. Assume there are m travelers with departure time t_i for $i = 1, \ldots, m$ in this period (i.e., $T_1 \leq t_i < T_2$). Then the accuracy measure is defined as the mean relative error within this period, i.e.,

$$E_{T_1,T_2} = \frac{\sum_{1 \leq i \leq m} e_i}{m}. \tag{7}$$

Here E_{T_1,T_2} is the accuracy measure defined on the time period and e_i is the relative error of the estimated travel time for the $i-th$ driver.

2.2.2 Relevance Measure

The *relevance measure* represents the variation of relative errors in an absolute sense. Mathematically, it is defined as the 75th percentile of the relative error's absolute values for all observed trip times in a given period. Therefore, the relevance captures both the accuracy and spread of the estimate. Denote R_{T_1,T_2} as the relevance measure for time period of $[T_1, T_2)$, we can define the relevance in such a way that:

$$Prob(|e| \leq R_{T_1,T_2}) \geq 0.75. \tag{8}$$

Here e is the relative error of an arbitrary vehicle traveling at a time during the period $[T_1, T_2)$. To better illustrate the concepts of accuracy and relevance measures proposed in this report, we give an example below. Table 1 lists the actual and estimated travel times for individual drivers for a given time period (in total 15 drivers). The relative error (calculated by equation (6) and the absolute value of the relative value for every driver are also shown in Table 1.

Driver Idx	Act. TT (min)	Est. TT (min)	Relative Error (%)	Absolute Value (%)
1	18'27"	16'57"	-8.10	8.10
2	18'58"	16'57"	-10.60	10.60
3	18'56"	16'57"	-10.44	10.44
4	18'30"	16'57"	-8.34	8.34
5	20'12"	16'57"	-16.06	16.06
6	20'13"	16'57"	-16.13	16.13
7	20'47"	19'45"	-4.93	4.93
8	20'23"	19'45"	-3.07	3.07
9	20'45"	19'45"	-4.78	4.78
10	21'25"	19'45"	-7.75	7.74
11	21'41"	19'45"	-8.88	8.88
12	21'24"	20'59"	-1.97	1.97
13	20'48"	20'59"	0.86	0.86
14	20'59"	20'59"	-0.02	0.02
15	21'13"	20'59"	-1.12	1.12

Table 1: Example of Quality Measures

We can then compute the accuracy and relevance measures. The accuracy measure is the mean

value of the forth column: -6.76%; the relevance is the 75-th percentile value of the fifth column: 8.80%.

3 Travel Times under Wave Propagation

This section presents the travel time estimation problem when a backward wave starts propagating. This type of situation arises whenever a *bottleneck* (BN) becomes active (i.e. free flow conditions downstream the BN and the presence of a queue upstream the BN). First, we will present the traffic flow model and how trajectories can be constructed. We will then study the performances of travel time estimates for given numbers of detectors.

3.1 Traffic Flow Model

The purpose of this section is to characterize the situation in which a BN becomes active and a backward-moving wave propagates in traffic. In particular, we are interested in identifying the parameters that characterize such a situation and propose some realistic values for them.

Empirical studies of a BN can be found in the literature. Studies of freeways located in Canada [16], [17], USA [18], and Europe (Germany) [19] have been reported. This report is mainly based on the conclusions of the work done by Cassidy and Bertini in [16] on two different freeways.

3.1.1 Traffic Characteristics under Wave Propagation

A bottleneck is said to be active if there is a queue upstream of it and free flow conditions downstream. Prior to the activation (and to the queue formation), the outflow at the bottleneck location is high. The duration of this high outflow is short and varies from day to day. Once the bottleneck becomes active, and a queue starts forming, the outflow drops dramatically at the bottleneck.

While the bottleneck is active, its outflow shows nearly-stationary patterns during the whole (congested) period. After the initial drop, it recovers again (the duration of the recovery also varies from day to day). Depending on the location, there might be more than one recovery periods. Even though these fluctuations are not reproducible from day to day, the long-run average outflow is. For that reason, some authors refer to this average outflow as the bottleneck capacity.

Once the cause of the bottleneck (for instance, high freeway demand, high on-ramp flows,

incident, etc.) is "removed", the outflow is high again until the situation is normalized.

In terms of the $q-k$ diagram at the bottleneck location x_0, and $x-t$ diagram, the situation is as shown in Figure 2. High outflow prior to the bottleneck activation corresponds to point A. Once the bottleneck becomes active, a backward-moving wave (a shock) propagates (with speed w_{AB}) and we have the following situation:

- Upstream the BN, between x_0 and the wave front: point B. (Upstream the wave front we are still at point A.)
- At the exact BN location, x_0: in theory, a new $q-k$ diagram can be used to represent what happens at x_0. In Figure 2, this corresponds to the smaller $q-k$ diagram (thicker line) and x_0 will be at point C (note that this is only a hypothetical point).
- Downstream the BN: point D.

Once the BN cause is "removed", a new backward-moving wave emanates from x_0 (with speed w_{BE}) and the situation is as follows:

- Upstream the BN, between x_0 and the second wave front: point E.
- Upstream the BN, in between both waves: point B. (Upstream the first wave front we are still at point A.)
- Downstream the BN: point E. After a while it will be point A again.

From the $q-k$ diagram, it can be seen that the second wave is always faster than the first one ($|w_{BE}| \geq |w_{AB}|$), so eventually it will catch up the first one. When that happens, it means that the queue has dissipated.

The above discussion is for a bottleneck caused by an incident. In practice, there are bottlenecks that are due to excessive demand. Figure 3 illustrates how traffic evolves under such condition. In the figure, traffic flow is low at the very beginning (at state F). Then it becomes higher than the bottleneck capacity (at state A) and stays constant for a period of time: t_A. After that, it recovers from A to F. The major difference between the incident case and excessive demand case is that in the latter case, the queue is dissipated via a forward (instead of backward) wave between states F and B.

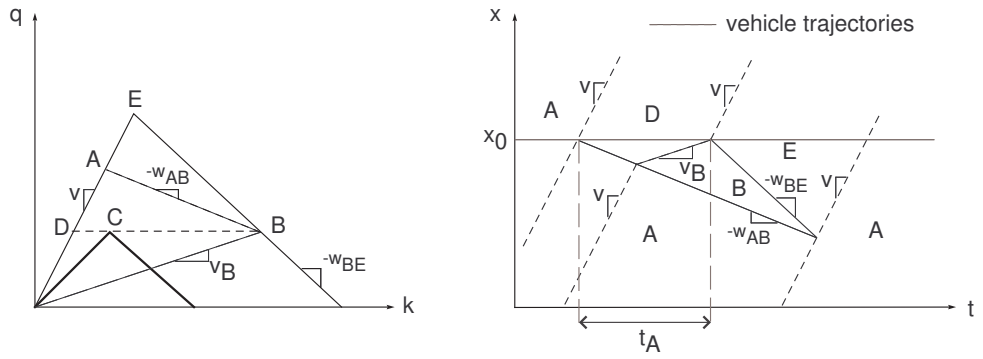


Figure 2: $q-k$ and $t-x$ diagram when backward-moving wave propagates (Incident).

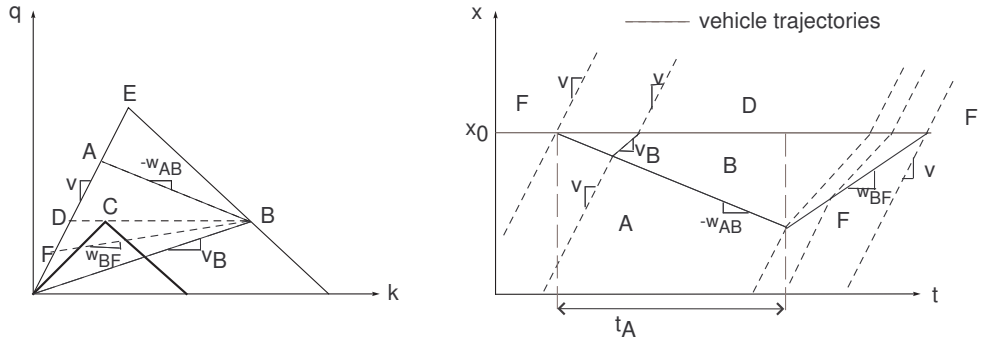


Figure 3: $q-k$ and $t-x$ diagram when backward-moving wave propagates (Excessive Demand).

3.1.2 Constructing Vehicle Trajectories

Vehicle trajectories can be analytically constructed under wave propagation as shown in Figures 2 and 3. For example, for the incident case, a vehicle will travel at speed v until it reaches the shockwave, as shown in Figure 2. At that point it reduces the speed to v_B until it reaches the second wave. After that, it start traveling at v again. The time that the vehicle travel at v_B can be computed using geometry. For instance, the maximum time that a particular vehicle travels at v_B is given by $t_{\max} = t_A \cdot \frac{w_{AB}}{v_B + w_{AB}}$. Similarly, trajectories for excessive demand cases can also be constructed as shown in Figure 3.

Therefore, if the starting location and time of a vehicle is given in the $x-t$ diagram, the trajectory of the vehicle can be constructed. For the incident case, the critical input parameters are:

- $q_{\max}, v_f, w, k_{\text{jam}}$: the capacity, free flow speed, wave propagation speed, and jam density of the freeway regular locations;
- x_b: the location of the bottleneck;
- C_b: capacity of the bottleneck;
- t_1, t_A: starting time and duration of the bottleneck activation; and,
- q_A: traffic flow.

For the excessive demand case, besides the parameters listed above, the traffic flow on state F is also required, denoted as q_F.

3.1.3 Realistic Values of Model Parameters

Besides the parameters of the fundamental diagram, points A, and B (also F if excessive demand case is considered) need to be determined in order to characterize the situation (remember that point C was only used to explain the theoretical condition at the exact BN location x_0, point D can be determined by knowing B, and point E depends on the parameters of the fundamental diagram). The time while the BN is active, t_A, is also needed to this end. Different combinations of these values will characterize the situation in terms of the length of the queue, how long vehicles have to travel at v_B, how many vehicle are affected by the BN, etc.

As it can be expected, the values of these parameters are site-specific. For instance, a BN might be activated for 30 minutes (an incident) or two hours (excessive demand), and flows depend on the number of lanes, etc. The drop in flow after the activation, seems to be about 8-10% for different cases, as was observed in [16]. The authors also provide some values for the flows before and after the activation (points A and B for incident case and points A , B and F for excessive demand case), and for the time duration.

Considering a 4-lanes freeway section, Table 2 shows three examples of typical parameters. The first two are for incident cases and the third one is for an excessive demand case. Parameters of the fundamental diagram are shown in Table 3.

Using the values in Table 2 and 3, the $t - x$ diagram can be constructed. The trajectories can then be used to evaluate travel time estimations given different loop configurations.

Case	A (vphpl)	B (vphpl)	F (vphpl)	t_A (hours)	v_B (mph)	w_{AB} (mph)	w_{BE} (mph)	w_{BF} (mph)
1	2000	1800	/	2	27.6	-6.3	-2.7	/
2	1800	1700	/	1.5	23.5	-2.4	-2.2	/
3	1750	1250	1000	0.5	12.0	-6.6	/	2.8

Table 2: Possible values for points A, B, and t_A, and the resulting values for v_B and w_{AB}.

Parameter	Value
Jam density: k_j	160 - 200 vpmpl
w	12-16 mph
Capacity: $q_{\max}$	2200 vphpl

Table 3: Parameters of the fundamental diagram for a 4-lanes freeway section.

3.2 Travel Time Performance for Wave Propagation

3.2.1 Experimental Design

To evaluate the performances of travel time methods, we select typical parameters for incident and excessive demand cases. As aforementioned, we deploy imaginary vehicles using certain fixed headway for a given period of time. This way we can obtain vehicle trajectories for a set of vehicles, which will also provide the actual travel times of all vehicles. For each vehicle, the estimated travel time can be calculated given a detector deployment scenario. The relative errors computed using the actual and estimated travel times of the vehicle can then be used to calculate the accuracy and relevance measures.

In this paper, we assume a vehicle headway of 2 seconds. For the incident case, we set $q_{\max}$ = 2200 vehicles/hour/lane, q_A = 1800 vehicles/hour/lane, C_b = 1600 vehicles/hour/lane, w = 14 miles/hour, v_f = 60 miles/hour, x_b = 2.5 miles, t_1=0.5 hour, and t_A = 0.4 hour. Therefore the bottleneck lasts for 24 minutes.

We further assume vehicles start traveling at $x_{\text{start}} = 0.25$ mile and end at $x_{\text{end}} = 2.55$ miles. We run the simulation for the period between 0.5 to 1.05 hours. Figure 4 illustrates the trajectories generated for the parameters listed above. Note that for ease of presentation, the vehicle headway was set as 30 seconds in the figure.

Vehicle trajectories are generated in a similar fashion for the excessive demand case. Besides the parameters for the incident case, we particularly set C_b = 1650 vehicles/hour/lane, q_F = 1500

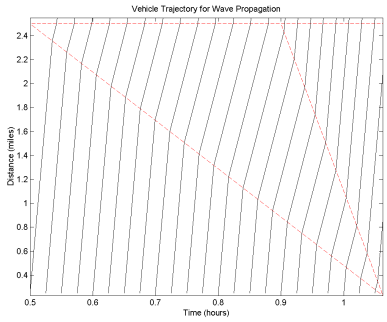


Figure 4: Trajectory for Incident Case

vehicle/hour/lane, and $t_A = 0.5$ hour. The starting and ending locations are set as $x_{\text{start}} = 0.95$ miles and $x_{\text{end}} = 2.55$ miles. Further, we set the simulation period from 0.5 to 1.5 hours. Figure 5 depicts vehicle trajectories generated using these parameters for the excessive demand case.

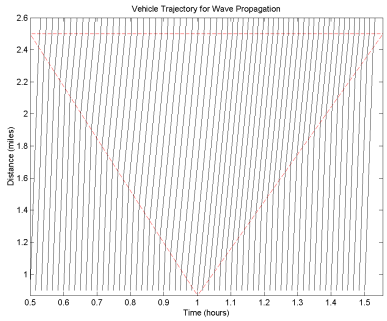


Figure 5: Trajectory for Excess Demand Case

In this report, we only test scenarios that sensors are evenly spaced. As aforementioned, a link is defined in such a way that its ending points are the middle points of two consecutive sensors. For the incident case, we vary the numbers of sensors from 1 to 23, resulting in detector spacing from 0.1 to 2.3 miles. While for the excessive demand case, we test the numbers of sensors from 1 to 16 which correspond to detector spacing of 1.6 to 0.1 mile. Furthermore, for multiple sensors, the results of instantaneous and dynamic travel times may be different. Therefore, performances for both methods will be evaluated.

3.2.2 Travel Time Performance for Incident Case

Figures 6 and 7 depict, respectively, the accuracy and relevance measures for the instantaneous and dynamic travel times for the incident case. The x-axis in these two figures is detector spacing

and y-axis is the relative error representing either accuracy or relevance.

One can observe that for both methods, the accuracy and relevance measures become improved as the detector spacing becomes smaller. For the instantaneous method, however, little improvement can be achieved when detector spacing is less than 0.5 mile. In this particular example, the accuracy measure stabilizes at 5% and the relevance fluctuates around 17% when the spacing is between 0.1 to 0.5 mile.

For the dynamic method, however, the performance can be continuously improved as spacing becomes smaller. When the spacing is 0.1 mile, the accuracy is near 0 and the relevance is less than 5%, which is significantly better than the instantaneous method.

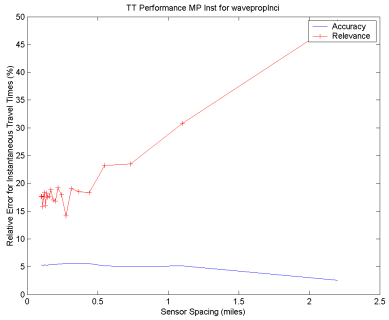


Figure 6: Performance of Instantaneous Travel Time (Incident Case)

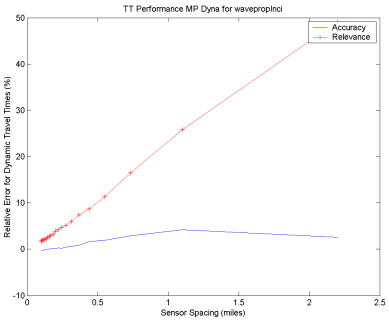


Figure 7: Performance of Dynamic Travel Time (Incident Case)

3.2.3 Travel Time Performance for Excessive Demand Case

Figures 8 and 9 depict, respectively, the performances (accuracy and relevance measures) for the instantaneous and dynamic travel time methods for the excessive demand case. We can observe that as the detector spacing decreases, the relevance keeps being improved. On the other hand,

the accuracy measure becomes slightly worse as the sensor spacing decreases. Further, there is no significant difference between the instantaneous and dynamic travel times for the excessive demand case.

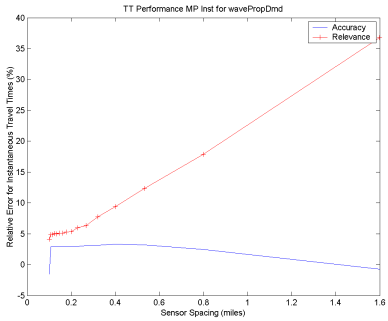


Figure 8: Performance of Instantaneous Travel Time (Excessive Demand Case)

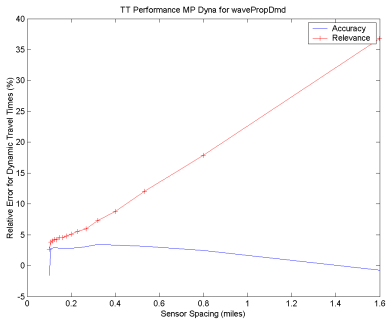


Figure 9: Performance of Dynamic Travel Time (Excessive Demand Case)

4 Travel Times under Oscillatory Traffic

Oscillations arise in queues that form upstream of an active bottleneck. We aim to characterize this oscillatory behavior, especially its speed profile, in order to provide a realistic travel time model for congested regimes.

4.1 Traffic Flow Model for Oscillatory Traffic

When a bottleneck becomes active, a queue forms upstream of it. Even though the outflow at the bottleneck is nearly constant, the flow in the upstream queue is probably not. The speeds of the vehicles in the queue oscillate between two values, a higher speed v_1 and a lower speed v_2. That is,

vehicles travel at certain speed v_1 for a period of time t_1, then decelerate to speed v_2 and maintain that speed for t_2, and finally accelerate again until speed v_1. The speeds v_1 and v_2 and the times t_1 and t_2 will probably change as the vehicle is traveling.

Based on these considerations, if we measure individual speed at a fixed location, we would find several vehicles circulating at v_1 for some time T_1 ($T_1 \neq t_1$), then a transition period with speed measurements between v_1 and v_2, and finally several other vehicles traveling at v_2 for certain time T_2 (where $T_2 \neq t_2$). The values of these four parameters (v_1, v_2, T_1, and T_2), and the transition between the two speeds fully characterize the oscillation at a given location.

4.1.1 Speed Profiles for Oscillatory Traffic

There are several ways of modeling the speed profile at a given location when oscillations start, depending on how the transition period (i.e. the period from decelerating from v_1 to v_2 or accelerating from v_2 to v_1) is modeled. The simplest approach would dismiss the transition period and consider a step function between the two speeds. This is shown in Figure 10. In this case, the four parameters (v_1, v_2, T_1, and T_2) need to be identified to characterize the oscillation. We refer this model as the "two-speed" model.

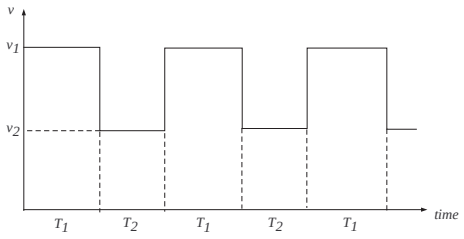


Figure 10: A simple speed profile.

There are other ways to model the transition period. For instance, transition between the two speeds might be just a line, or a curve, or an echelon function, as depicted in Figure 11 a, b, and c, respectively. In any of these three cases, more information –in addition to the four parameters previously mentioned– is needed to characterize the oscillations. For the first one, the slope of the line is needed (or the time that the transition period lasts), for the second one the equation of the curve is requested, and for the last one intermediate speeds and the times spent on each speed are needed. However, it is not clear how these extra parameters can be obtained in real life (no studies have previously attacked this problem).

More importantly, for the purpose of computing travel times, we argue that accurately capturing

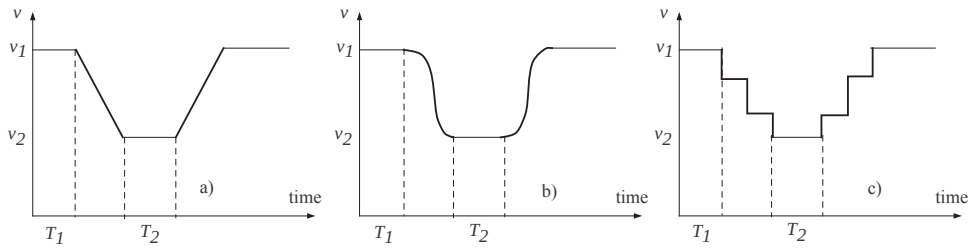


Figure 11: Different transition behaviors.

the detailed transition may not be necessary. For example, Figure 12 shows the speed profile of a particular vehicle under the oscillatory traffic (the thin line). The vehicle travels under speeds v_1 and v_2 for time t_1 and t_2 respectively. The transition between v_1 and v_2 takes arbitrary profiles and is usually complicated to model (thin line in Figure 12) . Nevertheless, to compute travel times, we are interested in the vehicle trajectory which is the "time integral" of the speed profile; in other words, we only need to focus on the area under the speed curve. As a result, the speed curve shown in Figure 12 can be approximated by the two-speed model and fully represented using four parameters (v_1, v_2, t_1', t_2'). Further, t_1' and t_2' are such that the two areas labeled with 1 are the same (and the two areas labeled with 2 are also the same). Since we do not know the real speed profile during the transition, we can assume $t_i' = t_i + \frac{t_{\text{tran}}}{2}$ for $i = 1, 2$, where t_{tran} is the transition time between v_1 and v_2. For short transitions (i.e. small t_{tran}), we expect that t_1' (t_2') will not deviate significantly from t_1 (t_2).

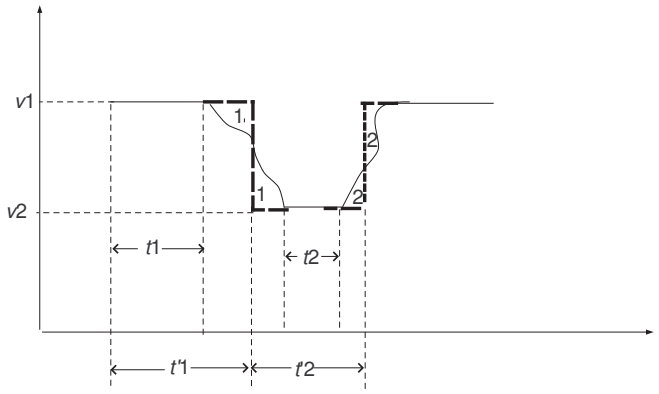


Figure 12: Characterizing Speed Profiles: speed profile of a particular vehicle.

In summary, we apply the two-speed model in this study to model the traffic under oscillatory conditions. Note that the simplified speed profiles are used only for computing travel times and may not be appropriate to study detailed traffic dynamics (e.g., transitions of traffic states). The

latter does need to consider detailed speed profiles, as shown in [20].

4.1.2 Constructing Vehicle Trajectories

Under the "two-speed" assumption, we can construct vehicle trajectories for oscillatory traffic provided the speed profile and fundamental diagram are known. This was first discussed by Coifman [21] and is illustrated in Figure 13. Basically, the two speeds represent two distinct traffic states, as marked "1" and "2" in the fundamental diagram in Figure 13b). The vehicle trajectory then oscillates between states "1" and "2", delineated by the shock wave with a backward speed $-w$. In other words, if the speed profile at a given location looks like that in Figure 10, the speed profile of vehicles will also be a step function between v_1 and v_2, but with different durations on each speed (t_1 for v_1, and t_2 for v_2).

Note that t_1 (or t_2), the duration in which a vehicle experiences v_1 (or v_2) is different as that recorded by the detector, i.e., T_1 (or T_2). In particular, the t's are related to the T's through the wave speed ($-w$, as shown in Figure 14). Mathematically, we have:

$$T_i = t_i \left(1 + \frac{v_i}{w}\right), i = 1, 2. \tag{9}$$

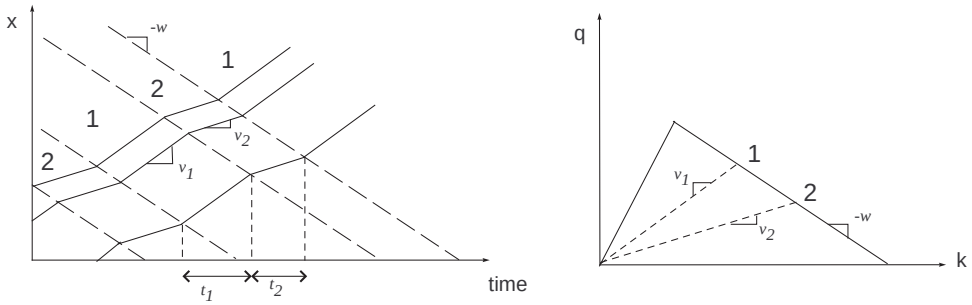


Figure 13: a) Trajectories from the two-speed model, b) Fundamental diagram.

It is important to note that when the fundamental diagram is triangular and the highway is congested, the above method to construct vehicle trajectories actually satisfies the LWR[2] continuity equation and therefore the Rankine-Hugoniot equation for shocks. That is, the trajectories that are constructed satisfy the kinematic wave theory based on the work of Lighthill and Whitham in [22] and of Richards in [23] (see Appendix A).

[2]Lighthill, Whitham, and Richards

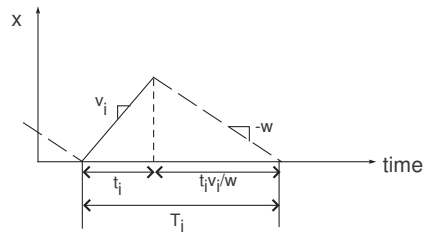


Figure 14: Relation between t_i and T_i.

4.1.3 Realistic Speed Profiles and Typical Values

So far we have shown that a two-speed profile (Figure 10) may provide a proper approximation of the oscillation phenomena. This section presents some realistic values for the parameters needed to characterize the oscillations, and it is based on information obtained from PeMS [24], and from other work shown in [25] and [26].

It has been shown that the values of v_1 and v_2 depend on the location in the queue with respect to the bottleneck: the further upstream of the bottleneck the location is, the lower the speeds v_1 and v_2 are. We have experimented with some scenarios with different values of v_1 and v_2. Table 4 shows three possible cases.

Location w.r.t. the bottleneck	v_1	v_2
Close	55	25
Less Close	30	10
Far away	10	0

Table 4: Possible values of v_1 and v_2 (in mph).

Regarding to the time parameters, T_1 and T_2 are in the order of minutes, and they can vary from 2 to 5 minutes. It is important to note that the method to reconstruct trajectories assumes that the oscillations are stable (i.e., they do not grow or dissipate over time and space).

With these values, we are able to characterize an oscillation. In order to simulate trajectories, however, we also need parameters to determine the fundamental diagram. These parameters are shown in Table 3.

4.2 Travel Time Performance for A Single Detector

In this section, we investigate the performance of estimated travel times under oscillatory traffic for the simplest scenario: a single detector case. Since we associate one link to a detector as mentioned in Section 2.1.2, the instantaneous and dynamic travel time estimates will be exactly the same, which can be computed using equation (3).

We conduct experimental studies in this section, applying typical parameter values as listed in Tables 4 and 3. We start with the experimental design in Section 4.2.1, followed by performances of estimated travel times vs. link lengths, speeds and time durations of the speeds. In this report, our primary interest in how the travel time estimation quality changes as the link length varies. However, for the single-detector case, we also show the sensitivity of travel time estimation quality vs. speeds and time durations of speeds in Section 4.2.3 and 4.2.4 respectively.

In Section 4.2.5, the limiting values of relative errors (defined in equation (6)) are analyzed and presented when the link length is very large. This analysis, on the one hand, illustrates the limitations of the two-speed model especially when the length of the link is relatively long. On the other hand, it also shows that the limiting errors of the two-speed model likely represents the worst-case scenarios if the actual speed profiles have more fluctuations within the limits of the two speeds.

4.2.1 Experimental Design

The two-speed model, which is represented by four parameters (v_1, v_2, T_1, T_2), is used in the experimental studies. Based on this model, the trajectories of individual vehicles can be simulated using the method proposed in [21] which is also described in Section 4.1.2.

First, we assume that the length of the link is L mile and a detector is located in the middle of the link, as shown in Figure 15. In this figure, the link is located from $x = 0$ to $x = 2$ mile and the detector is located at $x = 1$ mile. The figure also illustrates that, due to propagation of shock-waves, the time-space plane is divided into distinct regions, with speeds as v_1 and v_2 respectively. To simplify our discussion, we also assume at time $t = 0$, a shock wave just passes the entrance point of the subject link (i.e., the origin).

We further assume that the first vehicle enters the link at time $t = 0$. From the fundamental diagram, the entrance times of following vehicles to the link can be determined by the time headway h. To see this, we focus on Figure 13b. For a given speed v_1 (or v_2), the flow rate corresponding

to the speed can be uniquely determined provided that traffic is in the congested regime (which is the case for oscillatory traffic). Use v_1 as an example, the flow rate, q_1 can be computed as:

$$q_1 = \frac{w v_1 k_{\text{jam}}}{w + v_1}. \tag{10}$$

Then the time headway between consecutive vehicles in this speed region can be calculated by:

$$h_1 = 3600/q_1. \tag{11}$$

For speed region v_2, the time headway can be determined in a similar fashion. Both h_1 and h_2 are illustrated in Figure 15.

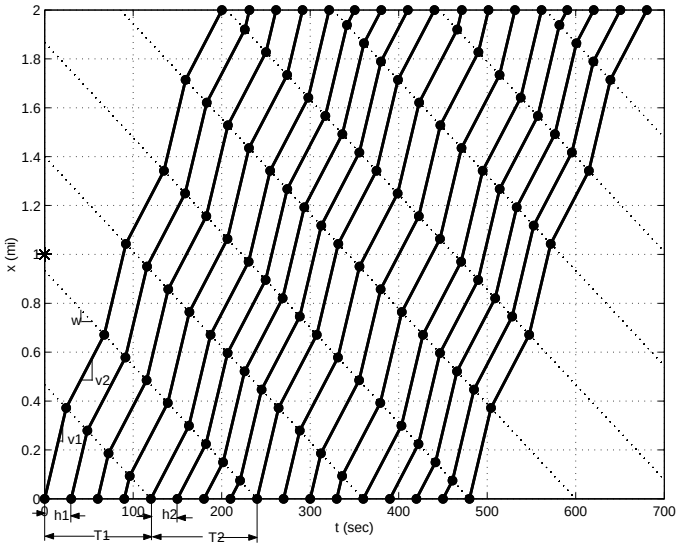


Figure 15: Experimental Design and Vehicle Trajectories

Using the above method, as shown in Figure 15, the trajectories of a group of vehicles can be determined for a study period, denoted as $[0, T]$. We assume there are N vehicles in this period (determined by the time headway). Given vehicle trajectories, the actual and estimated travel times of any vehicle can be calculated as in Section 2.1. In particular, we assume that the sample time period for computing average detector speed is set as $\Delta t = 30$ seconds (see Section 2.1.2). With individual vehicle's actual and estimated travel times in place, one can compute the relative error as described in Section 2.2.

In this study, we set the jam density $k_{\text{jam}} = 180$ vpmpl (vehicle-per-mile-per-lane), shock wave speed $w = 14$ mph, capacity $q_{\max} = 2200$ vph, and the simulation time $T = 3600$ seconds (1 hour).

Hence, for given parameters of the two-speed model (v_1, v_2, T_1, T_2) and a fixed link length L, one can compute the accuracy and relevance measures for the period $[0, T]$ as described above. In the next three subsections, we will investigate how the link length and the four parameters of the two-speed model impact the accuracy and relevance measures for a single link.

4.2.2 Travel Time Performance vs. Link Length

To study how link length impacts performances of travel time estimation for a single link, we vary the length of the link from 0.1 mile to 3 miles using 0.1 mile as the increment. For a given link length, we set $T_1 = T_2 = 120$ seconds, and test on the first two speed combinations listed in Table 4, i.e., ($v_1 = 55$ mph, $v_2 = 25$ mph) and ($v_1 = 30$ mph, $v_2 = 10$ mph). Figure 16 depicts how the accuracy and relevance measures vary as the link length changes by setting $v_1 = 55$ mph and $v_2 = 25$ mph. Figure 17 is for the speed combination of $v_1 = 30$ mph and $v_2 = 10$ mph.

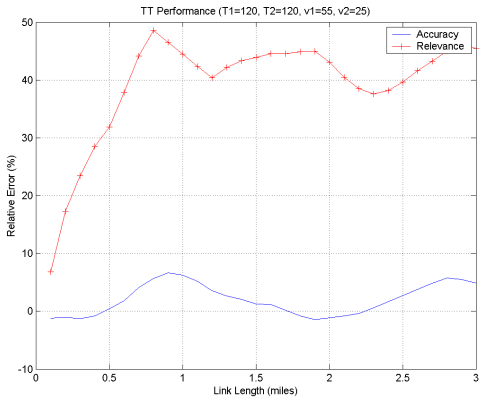


Figure 16: Accuracy and Relevance vs. Link Length ($v_1 = 55$ mph, $v_2 = 25$ mph)

The two figures depict that as the link length increases from 0.1 mile to 1 mile, both the accuracy and relevance measures degrade, while the relevance measure degrades more significantly. Further, if the link length is beyond 1 mile, the relevance becomes flat and the accuracy measure fluctuates within $\pm 10\%$. Therefore, the simulation results suggest that a detector should be deployed every one mile or less (preferably 0.5 mile or less). Note that this finding is only based on the two-speed model which, in most cases, may only provide a worst case scenario for evaluating detector spacing (see Section 4.2.5).

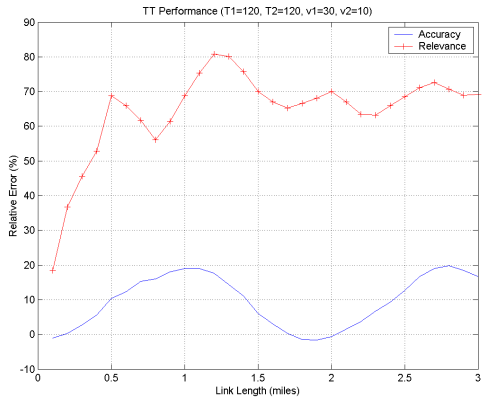


Figure 17: Accuracy and Relevance vs. Link Length ($v_1 = 30$ mph, $v_2 = 10$ mph)

4.2.3 Travel Time Performance vs. v_1 and v_2

To test the influence of v_1 and v_2, we fix the length of the link as 0.5 mile and $T_1 = T_2 = 120$ seconds. We then test for two scenarios. In the first scenario, we set $v_1 = 55$ mph and vary v_2 from 5 to v_1 using 5 mph as the increment. For the second scenario, we set $v_1 = 30$ mph and vary v_2 from 5 to v_1 using 5 mph as the increment. Figures 18 and 19 depict how the accuracy and relevance measures change with different v_1 and v_2 combinations respectively.

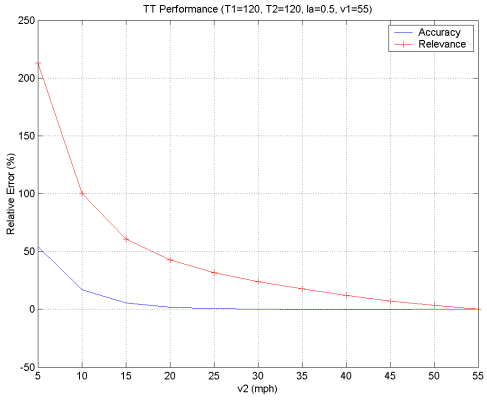


Figure 18: Accuracy and Relevance vs. v_2 ($v_1 = 55$ mph)

Both figures show a clear trend such that as the difference between v_1 and v_2 becomes smaller, the accuracy and relevance measures improve monotonically. This coincides with intuition. Since the detector can only capture either v_1 or v_2 at any given time instant in the two-speed model (except for the period that covers both speed regions), the error associated with this "imperfect"

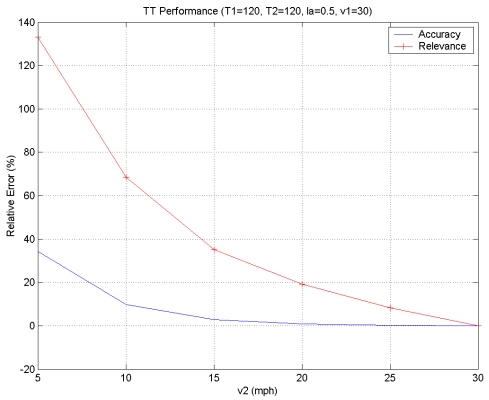


Figure 19: Accuracy and Relevance vs. v_2 ($v_1 = 30$ mph)

detection is expected to decrease if v_1 and v_2 are close to each other. In particular, if $v_1 = v_2$, the detector happens to sense "perfectly" the speed of the traffic state and thus the two measures will be zero, which is shown in Figures 18 and 19.

4.2.4 Travel Time Performance vs. T_1 and T_2

To study the impact of T_1 and T_2 on the travel time estimation performances, we fix the length of the link as 0.5 mile. We then vary T_1 from 30 to 210 seconds and $T_2 = 240 - T_1$ using 30 seconds as the increment. We then test on the first two speed combinations in Table 4. Figures 20 and 21 depict how the performance measures change with different T_1 and T_2 combinations.

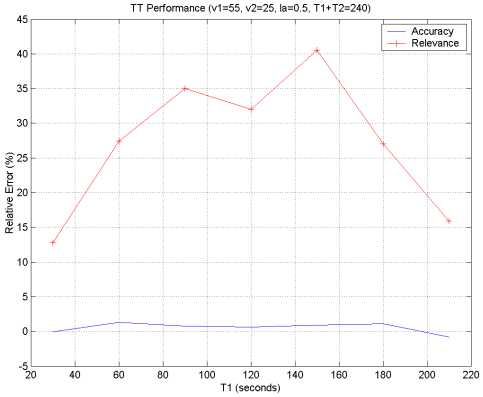


Figure 20: Accuracy and Relevance vs. T_1 ($v_1 = 55$ mph, $v_2 = 25$ mph, $T_2 = 240 - T_1$)

We observe from these two figures that the variation of T_1 and T_2 have less significant impact

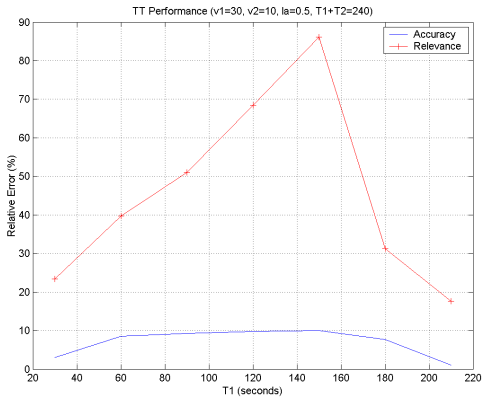


Figure 21: Accuracy and Relevance vs. T_1 ($v_1 = 30$ mph, $v_2 = 10$ mph, $T_2 = 240 - T_1$)

on the accuracy performance of travel time estimates compared with link length and traffic speeds. However, the relevance performance degrades substantially as T_1 and T_2 become closer (i.e., when T_1 is around 120 seconds since we set $T_1 + T_2 = 240$ seconds). This also coincides with intuition since if $T_1 \gg T_2$ (or likewise), v_1 (or v_2) will be dominant and the possibility of the detector captures the correct speed is expected to increase.

4.2.5 Relative Error of A Very Long Link

In this section, we show that as the link length becomes very large, the relative error of the estimated travel time (instantaneous travel time) can be computed analytically. Firstly, as the link length $L \to \infty$, the trajectory of a vehicle tends to cross the *same* number of high and low speed regions. Then the average speed of a vehicle traveling in oscillatory traffic can be computed as:

$$\bar{v} = (v_1 t_1 + v_2 t_2)/(t_1 + t_2), \tag{12}$$

where t_1 and t_2 are the times of the vehicles traveling in speed v_1 and v_2 respectively. According to equation (8), t_i can be represented by T_i as:

$$t_i = \frac{T_i w}{w + v_i}, \tag{13}$$

Substitute equation (13) into equation (12), one can obtain the average speed of a vehicle for traveling the entire link L as:

$$\bar{v} = L/t_r = \frac{T_2 v_2(v_1 + w) + T_1 v_1(v_2 + w)}{T_1(v_2 + w) + T_2(v_1 + w)} \tag{14}$$

Clearly, we should have $v_2 \leq \bar{v} \leq v_1$.

Since a detector can only detect either v_1 or v_2 at any time instant (note that we do not aggregate sensor speeds in this case), the estimated travel time from detector speed will be $\tau_{est} = L/v_1$ or $\tau_{est} = L/v_2$. Thus the relative error of any individual vehicle will be

$$e_1 = (\bar{v}/v_1 - 1) \tag{15}$$

$$e_2 = (\bar{v}/v_2 - 1) \tag{16}$$

Equations (15) and (16) provides two limiting values of the relative errors of estimated travel times when link length is infinite. Clearly, we have $e_1 \leq 0$ and $e_2 \geq 0$.

Figure 22 depicts that as link length becomes very large, the relative errors do converge to these two limiting values. In this figure, the two solid lines represents e_1 and e_2 respectively. We set parameters as follows: $T_1 = T_2 = 120$ seconds, $v_1 = 55$ mph, v_2=25mph, and L varies from 0.1 to 20 miles using 0.1 mile as the increment.

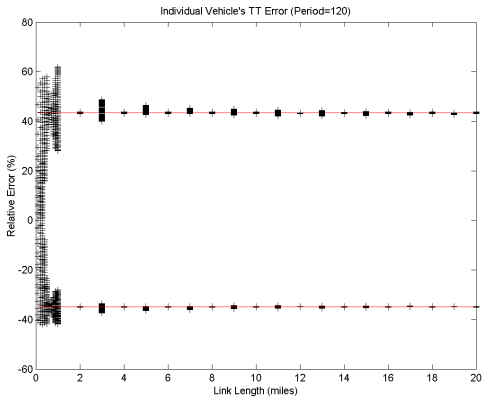


Figure 22: Limiting Relative Errors

This figure thus illustrates that the two-speed model may not be applicable for modeling very sparse sensors (i.e., associated links are very long). The above analysis also implies that the two limiting errors in equations (15) and (16) represent the maximum possible estimation errors if

certain transition exists between the two speeds v_1 and v_2, provided v_1 and v_2 are the observed largest and smallest speeds respectively. This can be seen from the following. First, from (12) and (13), it is clear that the average speed of the two speed model is the weighted average of the two speeds. The weight for each speed is $t_i = \frac{T_i w}{w+v_i}$. Similarly, if the speed profile contains $n > 2$ steps, i.e., we have speeds $v_1 \geq v_2 \ldots \geq v_n$, the average speed will be:

$$\bar{v} = \frac{\sum_{i=1}^{n} v_i \frac{T_i}{v_i+w}}{\sum_{i=1}^{n} \frac{T_i}{v_i+w}} \tag{17}$$

It is easy to see that we have n limiting errors in this case: $e_1, e_2, \ldots, e_n$, which can be expressed as:

$$e_i = \bar{v}/v_i - 1. \tag{18}$$

Clearly, e_1 and e_n are the two worst-case estimation errors in terms of under-estimating and over-estimating actual travel times respectively. Therefore, if the actual speed profiles contains more speed steps, the limiting errors computed using the two extreme speeds represents the lower and upper bounds of the possible estimation errors. This implies that the two-speed model proposed and studied in this paper actually corresponds to the worst case scenarios in terms of assessing the travel time estimation error. Or in other words, if the actual speed profiles are smoother (i.e., have more steps between the two speeds), it is very likely that the actual estimation errors will fall within these two bounds.

4.3 Travel Time Performance for Multiple Detectors

4.3.1 Experimental Design

The experimental design for the multiple-detector case is similar to that for a single detector. In particular, vehicle trajectories can be generated in the same way as discussed in Section 4.2.1. As aforementioned, we only test scenarios that detectors are evenly spaced in this report. Correspondingly, a link is defined as the segment between the middle points of consecutive detectors. To simplify our discussion, we also set the following parameters for the experiment: $v_1 = 55, v_2 = 25, T_1 = T_2 = 120$. The length of the route is set as 3 miles.

4.3.2 Travel Time Performance vs. Sensor Spacing

First notice that for multiple detectors, one can compute both the instantaneous and dynamic travel times as discussed in Section 2.1.2. Figure 23 first depicts the accuracy and relevance measures for instantaneous travel times as the detector spacing varies from 0.1 mile to 3 miles; Figure 24 shows the performance measures for dynamic travel times.

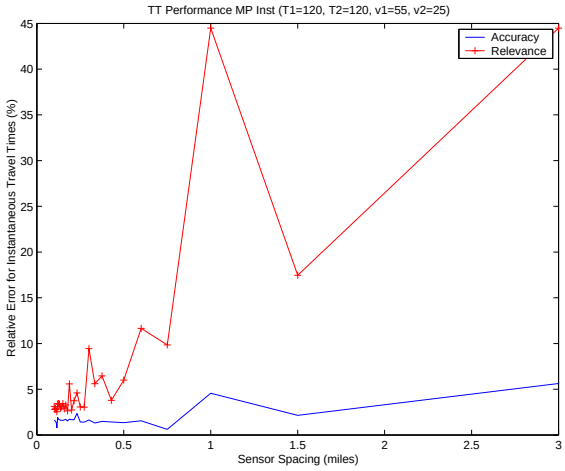


Figure 23: Performance of Instantaneous Travel Times

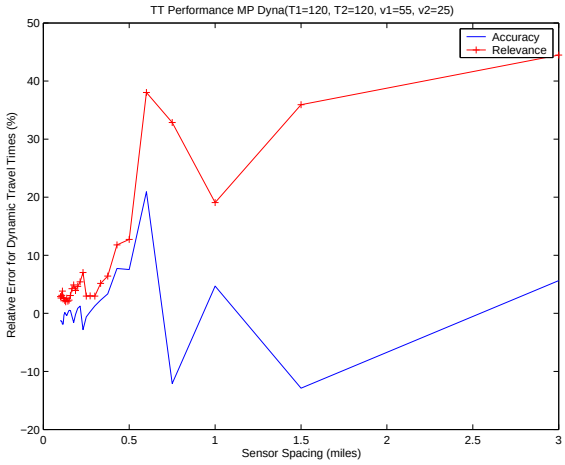


Figure 24: Performance of Dynamic Travel Times

We can observe from the two figures that as detector spacing becomes smaller, both measures especially the relevance measure are improved, although significant fluctuations may exist for both methods. These two figures also illustrate that when spacing is less than 0.5 mile, the accuracy and relevance measures are less than 10% for both the instantaneous and dynamic travel times.

Further, no significant improvement can be achieved for both methods if spacing is reduced from 0.5 mile to 0.1 mile. This suggests that 0.5 mile might be a reasonable detector spacing for providing travel times. At the 0.1 mile spacing, in particular, the accuracy and relevance measures are less than 5% for both methods.

5 Conclusion

We presented a simulation framework to study the required detector spacing for providing freeway travel times. The investigations were done by using simple yet widely used travel time estimation methods, namely the instantaneous and dynamic travel times. We first defined two quality measures to assess the travel time estimation quality: the accuracy measure and the relevance measure. The accuracy measure captures the mean of the travel time estimation error, while the relevance measure is defined on the variations of travel time estimation errors. Based on well-established traffic flow theory, simulation studies were conducted for two traffic conditions: wave propagation and oscillatory traffic. The former happens during incident occurrences, while the latter corresponds to traffic states under heavy congestion. For both conditions, we found that a sensor spacing of 0.5 mile is generally sufficient to obtain travel time estimates with errors less than 10%.

The work discussed in this report presents the first step of more rigorous investigations on optimal detector placement for travel time estimation and for other ITS applications in general. The findings in this report, e.g. a 0.5-mile spacing can generate travel time estimation errors less than 10%, may be site and model specific, and therefore may not be applied to other traffic conditions. For this reason, a more comprehensive framework that can capture various traffic conditions is needed. The authors have recently proposed a dynamic programming framework for optimal detector placement for freeway travel time estimation [27], which was also extended to derive optimal sensor placement for multiple ITS applications [28]. The results indicate that optimal sensor placement is highly correlated with bottleneck locations; with more sensors available, existing sensors that have already been deployed will remain unchanged and additional sensors will be deployed to bottlenecks to enhance existing sensors. More detailed discussions are provided in [27, 28].

References

[1] C. Chen, A. Skabardonis, and P. Varaiya. Travel time reliability as a measure of service. *Journal of Transportation Research Board*, 1855:74–79, 2003.

[2] H. Al-Deek and Emam B. Emam. New methodology for estimating reliability in transportation networks with degraded link capacities. *Journal of Intelligent Transportation Systems*, 10(3):117–129, 2006.

[3] A. Chen, H. Yang, H.K. Lo, and W. Tang. A capacity related reliability for transportation networks. *Journal of Advanced Transportation*, 33:183–200, 1999.

[4] Z.R. Peng, N. Guequierre, and J.C. Blakeman. Motorist response to arterial variable message signs. *Journal of Transportation Research Board*, 1899:55–63, 2004.

[5] H. Warita, T. Okada, and A. Tanaka. Evaluation of operation for travel time information on the metropolitan expressway. In *Proceedings of 8th ITS World Congess*, 2001.

[6] C.D.R. Lindveld, R. Thijs, P.H.L. Bovy, and N.J. Van der Zijpp. Evaluation of online travel time estimators and predictors. *Journal of Transportation Research Board*, 1719:45–53, 2000.

[7] Y. Iida, N. Uno, and T. Yamada. Experimental analysis approach to analyze dynamic route choice behavior of driver with travel time information. In *Proceedings of Vehicle Navigation and Information Systems*, pages 377–382, 1994.

[8] A. Khattak, A. Kanafani, and E.L. Colletter. Stated and reported route diversion behavior: implications of benefits of advanced traveler information system. *Transportation Research Record*, 1464:28–35, 1994.

[9] M. Kraan, N.V.D. Zijpp, B. Tutert, and et al. Evaluating networkwide effects of variable message signs in the netherlands. *Journal of Transportation Research Board*, 1689:60–67, 2000.

[10] J. Rice and E.V. Zwet. A simple and effective method for predicting travel times on freeways. In *Proceedings of IEEE Intelligent Transportation Systems*, pages 227–232, 2001.

[11] T. Oda. An algorithm for prediction of travel time using vehicle sensor data. In *Proceedings of Third International Conference on Road Traffic Control*, pages 40–44, 1990.

[12] X. Ban, Y. Li, A. Skabardonis, and J.D. Margulici. Performance evaluation of travel time methods for real time traffic applications. In *Proceedings of the 11th World Congress on Transport Research (CD-ROM)*, 2007.

[13] I. Fujito, R. Margiotta, W. Huang, and W.A. Perez. The effect of sensor spacing on performance measure calculations. In *Proceedings of the 85th Annual Meeting of Transportation Research Board (CD-ROM)*, 2006.

[14] http://www.ngsim.fhwa.dot.gov/.

[15] Berkeley Transportation Systems (BTS). Pems user guide, version 5.2. 2004.

[16] M.Cassidy and R.Bertini. Some traffic features at freeway bottlenecks. *Transportation Research B*, 33(1):25–42, 1999.

[17] M.Cassidy and M.Mauch. An observed traffic pattern in long freeway queues. *Transportation Research A*, 35(2):143–156, 2001.

[18] J.C. Munoz and C. Daganzo. Structure of the transition zone behind a queue. *Transportation Science*, 37(3):312–329, 2003.

[19] B.S. Kerner and H.Rehborn. Experimental properties of phase transition in traffic flow. *Physical Review Letter*, 79:4030–4033, 1997.

[20] B.S. Kerner. *The Physics of Traffic*. Springer, 2004.

[21] B.Coifman. Estimating travel times and vehicles trajectories on freeways using dual loop detectors. *Transportation Research A*, 36(4):351–364, 2002.

[22] M.J.Lighthill and G.B.Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. In *Proceedings of the Royal Society*, volume A 229, pages 317–345, 1955.

[23] P.I.Richards. Shock waves on the higway. *Operation Research*, 4:42–51, 1956.

[24] http://pems.eecs.berkeley.edu/public/.

[25] S.Ahn. *Formation and spatial evolution of traffic oscillations*. PhD thesis, University of California, Berkeley, Fall 2005.

[26] M.Mauch. *Study of oscillation in freeway traffic*. PhD thesis, Institute of Transportation Studies, University of California, Berkeley, 2001.

[27] X. Ban, R. Herring, JD Margulici, and A. Bayen. Optimal sensor placement for freeway travel time estimation. *Submitted to the 18th International Symposium on Transportation and Traffic Theory*, 2008.

[28] X. Ban, L. Chu, R. Herring, and JD Margulici. Optimal sensor placement for both traffic control and traveler information applications. In *Proceedings of the 88th Transportation Research Board Annual Meeting (CD − ROM)*, 2009.

Appendices

A Proof that Coifman's approach satisfies LWR equations

From the $q-k$ diagram (Fig.25), we know that $q = k \cdot v$ holds everywhere. In particular, it holds in the congested branch of the diagram, where $q = w \cdot (k_j - k)$. These two equations combined yield the following two equations:

$$\left\{ \begin{array}{rcl} q & = & \frac{v \cdot w \cdot k_j}{v+w} \\ k & = & \frac{w \cdot k_j}{v+w} \end{array} \right\} \tag{19}$$

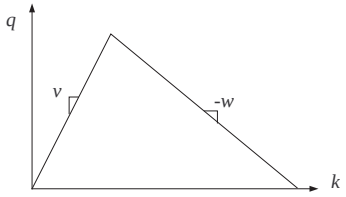


Figure 25: Fundamental diagram (triangular).

We will prove that equations in (19) satisfy the LWR continuity equation and the Rankine-Hugoniot equation. The first one is of the form $k_t + q_x(k) = 0$ and it is satisfied because our expressions are constant over time and space, and then their derivatives are zero.

The Rankine-Hugoniot condition says that the shockwave speed between states a and b is given by the slope of the cord between a and b in the fundamental diagram. That is, $\frac{q_a - q_b}{k_a - k_b}$. Using Equation (19), the numerator and denominator of previous expression are as follows:

$$q_a - q_b = \frac{w^2 k_j (v_a - v_b)}{(v_a + w)(v_b + w)} \tag{20}$$

$$k_a - k_b = \frac{w k_j (v_b - v_a)}{(v_a + w)(v_b + w)} \tag{21}$$

Finally, dividing Equation (20) by Equation (21) we find that the shockwave speed is $-w$, as the LWR theory states.

SECTION 4 – CORE STRATEGIES

X. BAN, R. HERRING, JD MARGULICI, A. BAYEN. “OPTIMAL SENSOR PLACEMENT FOR PROVIDING FREEWAY TRAVEL TIMES”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2008.

H. LIU, A. DANCZYK. “OPTIMAL DETECTOR PLACEMENT FOR FREEWAY BOTTLENECK IDENTIFICATION”. CONFERENCE PAPER, 87TH ANNUAL TRANSPORTATION RESEARCH BOARD MEETING, 2008.

Optimal Sensor Locations for Providing Freeway Travel Times

Xuegang (Jeff) Ban * Ryan Herring† JD Margulici‡
Alexandre M. Bayen§

Abstract

This article presents a dynamic programming algorithm for determining optimal sensor locations used to compute freeway travel times. This is done by modeling the problem as a staged process, which leads to a formal dynamic programming framework. We show that a graph representation exists for the dynamic programming formulation and prove that a polynomial algorithm exists to solve the optimal sensor placement problem. Numerical examples are provided to illustrate the model and algorithm using microscopic traffic simulation data and GPS data from the Mobile Century experiment conducted by University of California, Berkeley, Nokia and California Department of Transportation (Caltrans).

1 Introduction

Intelligent Transportation Systems (ITS) applications rely on various types of data (such as traffic flow, speed, or occupancy), which are usually collected from traffic sensors. For example, freeway travel time estimation often requires speeds measured at certain locations. Traditionally, many traffic sensors were deployed on a case by case basis by practitioners without a systematic study of the quantity and locations of sensors needed[1]. Since traffic sensors are limited resources, determining optimal deployment strategies maximizes the value of that resource.

In this article, we study the optimal sensor placement problem for providing freeway travel times. The travel time application is selected because travel time estimates are one of the most useful roadway traffic metrics for both traffic management agencies and the driving public. First, travel time is a crucial and direct measure of traffic conditions and system performance. Travel time reliability has been receiving particular attention from the Federal Highway Administration (FHWA) and recent research aims to quantify travel time reliability at a network level [3], [4], [5].

*Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, banx@rpi.edu
†Department of Industrial Engineering and Operations Research, University of California, Berkeley, ryanherring@berkeley.edu
‡California Center for Innovative Transportation, University of California, Berkeley, jd@calccit.org
§Department of Civil and Environmental Engineering, Systems Engineering, University of California, Berkeley, bayen@berkeley.edu

[1]One exception is the optimal sensor location problem for origin-destination matrix estimation, which has been widely studied in the literature [1, 2]

Second, travel times represent information that is easy to understand and process. Numerous studies [6, 7, 8] reveal that commuters value travel time information, which reduces their uncertainty and stress. Furthermore, relevant information can arguably enable travelers to make educated choices about their itinerary, departure time or even transportation mode, which may result in a form of "system self-management."

In the past, numerous studies have contributed to algorithmic techniques to estimate travel times from available field data, which generally came from loop detectors [9, 10]. Most studies assumed given detector locations and proposed optimal ways of processing the data. The optimal sensor placement problem in this regard has not been widely studied. A few existing research efforts focused on empirical investigations of the impact of sensor locations on the quality of travel time estimation. Ozbay et al. [11] studied travel time estimation quality vs. sensor locations under both recurrent and non-recurrent congestion. By taking out existing loop detectors in a predefined way, Fujito et al. [12] found that travel time estimation quality (measured by the travel time index) may not always decrease as detector spacing increases. For a 9-mile route in California, Kwon et al. [13] studied how the travel time estimates vary as the number of detectors changes by randomly taking out detectors. They concluded that 0.5 mile sensor spacing is appropriate for providing freeway travel times. Ban et al. [14] showed that as sensor spacing increases, travel time estimation becomes more sensitive to actual sensor locations. This implies that the optimal sensor location problem is more critical if one needs to deploy a limited number of sensors on a relatively long freeway segment. The above studies are usually available for freeways. Thomas [15] and Oh et al. [16], on the other hand, studied sensor location problems for arterial streets using microscopic traffic simulation.

Relatively little research has been devoted to the development of tractable methods for optimal sensor placement for travel time estimation. Eisenman et al. [17] provide an information learning based conceptual framework of the sensor location problem for traffic detection systems. Sherali et al. [18] propose a mixed-integer optimization model to determine optimal placement of vehicle identification readers for travel time estimation, although the model can only be solved approximately. Bartin et al. [19] show that the optimal sensor placement for travel time estimation can be determined by minimizing the weighted summation of speed variations of all roadway segments, each of which is associated with a sensor. A nearest neighbor (NN) algorithm was further developed in [19]. However, the NN algorithm is not guaranteed to provide a globally optimal solution in polynomial time.

We believe that two major issues remain unresolved in terms of optimal sensor placement for travel time estimation:

(1) There is no existing model and algorithm that can efficiently solve (in polynomial time) the optimal sensor placement problem for freeway travel time estimation.

(2) In reality, most corridors are already equipped with some sensors. However, there is no algorithm that can provide optimal placement of additional sensors to supplement the existing sensors in polynomial time.

In this article, we focus on the above two issues. In particular, by discretizing both the time and space, we show that the optimal sensor location problem for travel time estimation can be formulated as a Dynamic Programming (DP) model with a sensor deployed at each stage. The model can be further represented as an acyclic graph and solving the problem is equivalent to finding the shortest path in the graph. We then prove that such a search can be done in polynomial time, which can be used for solving large scale problems or for deploying sensors to many freeway segments that need to be monitored. We also show that incorporating sensors that have already

been deployed can be easily done via revising the graph representation of the DP model, and the complexity of solving the model remains the same.

Distinct from most previous studies, we test the model and algorithm using both simulation data and GPS-equipped cellular phones. The results show that to have better travel time estimation, sensors should be deployed to cover major bottleneck areas and free-flow regimes. As more sensors are available, they should be placed at bottlenecks areas, while a single sensor is usually sufficient for free-flow areas.

The remainder of this article is organized as follows. We first give in Section 2 the formal definition of the problem, including the travel time estimation methods we use to illustrate the model, the sensor data that we use, and the objective function to be optimized. This section sets up the stage for developing the DP model in this article. Section 3 presents how the dynamic programming formulation is derived. We first discretize space into small *sections* and time into *intervals*. We show that if vehicle trajectories are available, the average speeds of each section at any time interval can be calculated or estimated. This results in the speed field of the study route for a given time period. By exploiting the speed field and available vehicle trajectories, we show that he optimal sensor locations can be determined in a staged process with one sensor deployed in a stage. This leads formally to a DP model. A recursive formulation for the DP model is also provided, together with the constraints and state transfer equations. In Section 4, a graph representation of the model is given. We show that the graph is acyclic and solving the DP model is equivalent to finding the shortest path in the graph, implying the solution complexity is polynomial. We also present in this section how to incorporate existing sensors into the DP model. We test the model and solution algorithm in Section 5 using both micro-simulation data and real-world data from GPS-based cellular phones. Section 6 concludes our work and provides discussions for future research.

2 Preliminaries

The problem studied in this article can be stated as follows: given a freeway segment (called *route* r) and a given number of fixed-location sensors (such as loop detectors), where should these sensors be placed so that their deployment is "optimal" in terms of providing travel time estimates? Here we assume the number of sensors is given (denoted as K), which may often be determined by budget constraints. If this is not the case, one can always solve the problem for different numbers of sensors and pick the one with the desired performance. The efficiency of our proposed algorithm in this article makes solving the problem multiple times (with different values of K) tractable.

Similar to other engineering problems, the answer to this optimal sensor placement problem depends on several factors. First, there are numerous methods available to compute travel times and sensors can usually provide multiple types of data (such as aggregated and disaggregated speeds, flow, occupancy, etc). Therefore, determining optimal sensor locations is dependent upon the travel time estimation method and the sensor data type. It also depends on other factors such as geometry of the freeway segment, bottleneck locations, travel demand, etc. This section discusses the assumptions used in the article to address these concerns, most of which are consistent with what is currently used in practice.

2.1 Travel Time Estimation Methods

To be consistent with current practice, we assume travel times are calculated based on aggregated sensor speeds (say every $\Delta T = 30$ seconds). Speeds can be obtained directly from double loop detectors and other types of fixed location sensors or estimated from single loop detectors [20]. We further assume that every sensor has a spatial "influence area", called a *link*. And the sensor speed represents the (uniform) speed of the entire link associated with the sensor. There are a number of ways that how a sensor is associated with its link (e.g., PeMS defines a link as the segment between the middle points of two sensors. See [21], pp. 3-1). In this article, we assume a sensor is always in the middle of its corresponding link [2]. It is our understanding that different link definitions are just (slightly) different ways to utilize sensor speeds and the travel time calculation results should not deviate too much.

Following this convention, the to-be-deployed K sensors divide the study route r into K links, and the route travel time is the summation of all link travel times. We recognize that such a definition will effectively eliminate certain travel time estimation methods based directly on routes (e.g., [9]). However, we notice that it is a widely used route travel time calculation method in practice (see for example [21], pp. 3-23). More importantly, the DP model presented in this article does not depend on how link travel times are calculated. This implies much flexibility regarding which travel time method to use in the model.

We focus on two specific travel time calculation methods in this article: the *instantaneous* and *Coifman* methods. The instantaneous method assumes traffic conditions remain unchanged from the time a vehicle enters a route until it leaves the route. Therefore, the travel time of the route can be computed by summing the travel times of the constituent links at the time a vehicle enters the route. This method is "naive" in the sense that traffic condition changes are not considered at all; however, it is probably the mostly widely used method in practice due to its simplicity and the fact that it can be used in real time (i.e., no future information or prediction is required). The second method, originally developed by Coifman [10], is a more sophisticated algorithm for calculating link travel times based on sensor speeds. The method estimates vehicle trajectories from sensor speeds by basic traffic flow theory, from which link travel times can be extrapolated. The reader is referred to [10] for detailed discussions of how the algorithm works.

We use the instantaneous method in most parts of the article to illustrate the DP model and the solution algorithm. However, we discuss how Coifman's method can also be considered in the model and solution method. In Section 5, we show results from the Coifman's method, and provide comparisons with those by the instantaneous method.

2.2 The Objective Function

This subsection defines the objective function that needs to be optimized in the DP model. For this purpose, we assume that we have trajectories of a certain number of vehicles (assumed to be M). We first denote $\hat{\tau}_k^m$ and τ_k^m the estimated and actual travel time of the m-th vehicle ($1 \le m \le M$) traveling link k ($1 \le k \le K$), respectively. Then the travel time estimation error for the m-th vehicle on link k, denoted as e_k^m, can be expressed as:

[2] One may argue that restricting sensors to be only in the middle of its link may potentially filter out better solutions. This is true from a pure optimization point of view and in fact was considered by some researchers. For example, a probability distribution was assumed in [19], which describes the probability that a sensor will be deployed to each discretized section of a link . However, in practice, after sensor are deployed based on the results from specific optimization models, practitioners need a straightforward way to define the link associated with each sensor to compute travel times. If sensors are allowed to be deployed at any arbitrary location within a link, the link boundary will have to be recorded to compute travel times. We argue that this is highly impractical in reality.

$$e_k^m = \hat{\tau}_k^m - \tau_k^m. \tag{1}$$

In this article, we use the same objective function as that in [19], which is defined as follows:

$$\hat{E} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K}(e_k^m)^2}{M} = \sum_{k=1}^{K}\hat{E}_k. \tag{2}$$

Here $\hat{E}$ represents the objective function. $\hat{E}_k$ is the Mean Square Error (MSE) of the travel time estimation for all M vehicles for link k, defined as:

$$\hat{E}_k = \frac{\sum_{m=1}^{M}(e_k^m)^2}{M}. \tag{3}$$

The objective defined in (2) focuses on estimation errors of all individual links, instead of only on the entire route. The reason for this is that we want to generate sensor locations that can provide "good" estimation for all link travel times, not only in terms of the entire route. If attention is only put on the entire route, it is possible that the resulting sensor locations may underestimate travel times for certain links and overestimate for other links, but as a whole, they cancel out each other and provide good estimation. This type of sensor placement is not desirable. It is easy to see that the objective function we use here can effectively eliminate such sensor deployment strategies since they will have large objective values using equation (2). Hence, we need to deploy sensors to minimize $\hat{E}$.

3 A Dynamic Programming Formulation

This section presents how the DP model can be derived. It starts with a discussion of how the link MSE can be calculated.

3.1 Mean Square Error of A Link

Since the objective here is to minimize the summation of link MSEs, we investigate the MSE of any link k. For this purpose, we apply a scheme to discretize both time and space. We first divide the given route r into small segments, called *sections*. The premise is that if the length of a section is sufficiently small, we can reasonably assume that speed does not change within the section and it does not matter where to place a sensor within the given section. Then we only need to determine where to deploy the given K sensors to these small sections. Assume the length of each section is Δx and that the given route r can be divided into N sections. We use $n = 1, \ldots, N$ to index a given section. A link then contains one or more sections, and the link boundaries are at the section boundaries. Also, since we assume a sensor is always in the middle of its link, the sensor deployment problem is now converted to determine the optimal starting and ending indices of all the K links that comprise the study route. In the time domain, it is natural to divide (evenly) the time into *intervals* with the interval length $\Delta T = 30$ seconds because we assume that sensors can only provide 30-sec average speeds. In particular, assume the entire study period can be divided into H time intervals and $h = 1, \ldots, H$ is used to index a given interval. In this article, assume route r starts with $x = 0$ and time starts with $t = 0$.

The idea of discretizing both space and time is illustrated in Figure 1. It is clear from the figure that the two-dimensional $x-t$ space is divided into grids, defined as *sensor boxes*. Each sensor box represents a data collection unit (particularly for speeds in this article) at a specific location (section), which is only active for the designated time period (30-sec long). The average speed of each sensor box can be computed via available vehicle trajectories. In particular, it is defined as the average speed of all vehicles that pass the sensor at the designated time period. Clearly, this mimics the way how loop detectors collect 30-sec average speeds in reality. Calculating the average speed for any sensor box $(n,h), \forall n = 1,\ldots,N, h = 1,\ldots,H$ within the route will result in the *speed field* (also called *speed contour map*, see [22]) of the study route for the study period. Figure 6(a) depicts the speed field of the micro-simulation data studied in this article.

Notice that if all vehicle trajectories are available and can cover all sensor boxes, the speed field can be calculated; otherwise, we may have "blank" sensor boxes for which there is no vehicle passing by. For those blank sensor boxes, we estimate their average speeds using surrounding sensor boxes, which is called *imputation* [23]. In this article, we adopt a simple imputation method: the speed of a blank sensor box is the average speed of all its surrounding sensor boxes whose speeds are already available. Figure 12 (b) illustrates the estimated speed field using trajectories from 100 GPS-equipped vehicles.

Assume the speed field is given by the above discretization scheme. Further assume the k-th link starts with section s_k and ends with section $y_k \geq s_k$. Both s_k and y_k are integers to represent a section. Note that the starting and ending sections are both inclusive, i.e., link k starts at the starting location of section s_k and ends at the ending location of section y_k. This is shown in Figure 1.

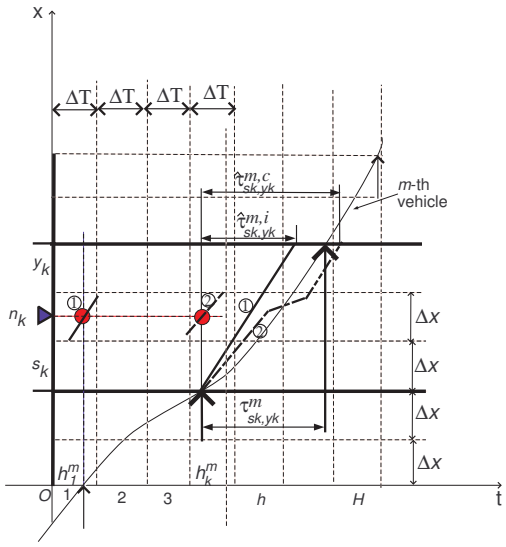


Figure 1: Actual and Estimated Travel Times.

To calculate the MSE of Link k as expressed in equation (3), we focus on the given M vehicles. For any $m-th$ vehicle, Figure 1 depicts, in a solid thin line, the trajectory of the vehicle. Then it is obvious that the actual travel time of the vehicle traversing link k is:

$$\tau^m_{s_k,y_k} = t^m_{y_k\Delta x} - t^m_{(s_k-1)\Delta x}. \tag{4}$$

Here t_x^m denotes the time when the m-th vehicle passes location x, and τ_{s_k,y_k}^m is the travel time of the m-th vehicle from the starting location of section s_k to the ending location of section y_k.

Suppose a sensor is deployed on this k-th link. Based on our assumptions in this article, a sensor will be in the middle of a link. Denote n_k the section that the sensor on Link k is located, we have:

$$n_k = \lfloor (s_k + y_k)/2 \rfloor. \tag{5}$$

Here $\lfloor \cdot \rfloor$ denotes the *rounding* operator. Assume the m-th vehicle enters route r at time interval h_1^m and it enters section s_k, the starting section of Link k, at time interval h_k^m. Then according to the definitions of the instantaneous travel time, the average speed of sensor box (n_k, h_1^m), denoted as v_{n_k,h_1^m}, will be used for computing the instantaneous travel time of Link k. This is shown as the solid bold line in Figure 1, which is marked as "1". Denote $\hat{\tau}_{s_k,y_k}^{m,i}$ the instantaneous travel time of the m-th vehicle traversing Link k. Also noticing that $(y_k - s_k + 1)\Delta x$ is the length of Link k, we thus have:

$$\hat{\tau}_{s_k,y_k}^{m,i} = \frac{(y_k - s_k + 1)\Delta x}{v_{n_k,h_1^m}}, \tag{6}$$

If Coifman's method is used instead for the link travel time, the vehicle trajectory will be first estimated by a piece-wise linear curve via basic traffic flow theory [10]. This is shown as the bold dash line in Figure 1 (also marked as "2"). The Coifman link travel time for link k, denoted as $\hat{\tau}_{s_k,y_k}^{m,c}$, does not have a close-form expression. However, it is clear that $\hat{\tau}_{s_k,y_k}^{m,c}$ only depends on the starting and ending sections of Link k provided speeds of all sensor boxes are given, and the entrance time is assumed to be $t_{(s_k-1)\Delta x}^m$.

Denote $\hat{E}_k^i$ and $\hat{E}_k^c$ the MSE of travel time estimation for link k for instantaneous and Coifman travel times, respectively. Following equation (3), they are both functions of (s_k, y_k) and can be expressed as:

$$\hat{E}_k^i(s_k, y_k) = \frac{\sum_{m=1}^M \left(\hat{\tau}_{s_k,y_k}^{m,i} - \tau_{s_k,y_k}^m\right)^2}{M} = \frac{\sum_{m=1}^M \left(\frac{(y_k - s_k + 1)\Delta x}{v_{n_k,h_1^m}} - t_{y_k\Delta x}^m + t_{(s_k-1)\Delta x}^m\right)^2}{M}, \tag{7}$$

$$\hat{E}_k^c(s_k, y_k) = \frac{\sum_{m=1}^M \left(\hat{\tau}_{s_k,y_k}^{m,c} - \tau_{s_k,y_k}^m\right)^2}{M}. \tag{8}$$

The above procedures for calculating link MSE (instantaneous or Coifman) show that the link MSE only depends on the starting and ending sections of the link, i.e., s_k and y_k. In particular, the calculation is independent of how the $(k-1)$ sensors for the previous $(k-1)$ links are deployed once s_k and y_k are known. This motivates us to formulate the optimal sensor placement problem using dynamic programming, as will be shown in the next section.

3.2 Dynamic Programming Model

Since we assume a sensor is in the middle of its associated link, the optimal sensor location problem can be solved via finding the optimal starting and ending locations (i.e., section numbers) of the

links. The objective is to minimize the function defined in (2). Denote $\hat{E}^i$ and $\hat{E}^c$ the objective function for instantaneous and Coifman travel times respectively. We will have, according to (2):

$$\hat{E}^i = \sum_{k=1}^{K} \hat{E}_k^i(s_k, y_k), \tag{9}$$

$$\hat{E}^c = \sum_{k=1}^{K} \hat{E}_k^c(s_k, y_k). \tag{10}$$

We can see that the objective functions for instantaneous and Coifman travel times are very similar, and the only difference is which link MSE to use. Therefore, in the reminder of this article, we only use instantaneous travel time to illustrate the proposed models and solution algorithms.

Given the objective function, the optimal sensor location problem can be stated as follows: find the optimal values of $(s_k, y_k), \forall k = 1, \ldots, K$ such that (9) can be minimized. That is, one needs to solve the following optimization problem:

$$\min_{1 \le s_k, y_k \le N, \forall k=1,\ldots,K} \quad \sum_{k=1}^{K} \hat{E}_k^i(s_k, y_k). \tag{11}$$

Subject to constraints (12) - (15) below.

The above optimization model is a linear integer program since $\hat{E}_k^i(s_k, y_k)$ is computable for any given (k, s_k, y_k), and (s_k, y_k) are integer-valued, $\forall k = 1, \ldots, K$. However, directly solving the model may not be easy if the dimension of the problem is large.

In this article, we divide the problem into stages: at each stage, the optimal location of a sensor is obtained, which can be achieved by finding the optimal starting and ending locations of its associated link. For that purpose, we denote the starting location (section) of link k (i.e., s_k) as the state variable. Accordingly, the ending location of link k (i.e., y_k) is the decision variable. Once s_k and y_k are given, the sensor location (section) can be achieved through equation (5).

We first look at the constraints for s_k and y_k. Clearly, we have

$$s_1 = 1, \tag{12}$$

$$y_K = N. \tag{13}$$

That is to say, the first link must start at Section 1 and the last link (Link K) must end at Section N. Also, we have the state transfer function as

$$s_{k+1} = y_k + 1. \tag{14}$$

That is, knowing the ending section of Link k (y_k), the starting section of Link $(k+1)$ must be the next section $(y_k + 1)$.

Further, since one link contains at least one section, we have

$$k \le s_k \le y_k \le N - K + k. \tag{15}$$

The first inequality holds since there are $k-1$ links before Link k, which contain at least $k-1$ sections. Similarly, the last inequality holds since there are $K-k$ links after Link k, which contain at least $K-k$ sections. Furthermore, equations (12) - (15) show that there are only one possible state for Stage 1 as $s_1 = 1$, but multiple states for Stage $k \geq 2$. In particular, equation (15) means that the possible states for Stage $k \geq 2$ is from k to $N-K+k$, i.e., the total number of states is $N-K+1$.

At any stage k, the cost of deploying a sensor is assumed to be $\hat{E}_k^i$, which is consistent with the objective function (9) and (2). Since $\hat{E}_k^i$ is only a function of (s_k, y_k), the optimal value of y_k can be obtained by minimizing $\hat{E}_k^i$ if s_k is known. In particular, if we denote $F_k(s_k)$ as the total cost from stage k (including stage k) to the last stage (i.e. stage K), a recursive formulation for $F_k(s_k)$ can be given as:

$$F_1(s_1) = F_1(1) = \min_{1 \leq y_1 \leq N-K+1} \left\{ \hat{E}_1^i(1, y_1) + F_2(y_1+1) \right\}, \tag{16}$$

$$F_k(s_k) = \min_{s_k \leq y_k \leq N-K+k} \left\{ \hat{E}_k^i(s_k, y_k) + F_{k+1}(y_k+1) \right\}, \forall 2 \leq k \leq K-1, \tag{17}$$

$$F_K(s_K) = \hat{E}_k^i(s_k, N). \tag{18}$$

The above three equations are for stage 1, stage $2 \leq k \leq K-1$, and stage K respectively. First, due to (12), we have $F_1(s_1) = F_1(1)$ for stage 1, which is a summation of the cost of stage 1 (i.e. $\hat{E}_1^i(1, y_1)$) and that from stage 2 to stage K (i.e. F_2). For stage $2 \leq k \leq K-1$, the cost F_k is a function of the state variable s_k, which is also the summation of the cost of the current stage k and that from stage $k+1$ to the last stage. Note that in both equations, the starting location of the next stage (i.e. stage 2 or $k+1$) is the *immediate next* section of the ending location of current stage (i.e. y_1+1 and y_k+1 respectively) due to (14). For the last stage, since the ending location must be N, $F_K(s_K)$ is automatically computable given s_K.

We can easily observe that 1) all constraints (12) - (15) are satisfied in the above three equations and there are no extra constraints introduced, 2) $F_1(1) = \sum_{k=1}^{K} \hat{E}_k^i(s_k, y_k)$. Therefore, solving (11) is equivalent to solve (16) - (18). Further, from these recursive equations, we can see that if $(s_k^*, y_k^*), 1 \leq k \leq K$ is an optimal solution, $(s_k^*, y_k^*), k_1 \leq k \leq k_2$ must be an optimal solution from stage $k_1 \geq 1$ to stage $k_2 \leq K$ [3]. This illustrates that the *optimality principal* [24] holds for the model (16) - (18). Therefore, the model is a Dynamic Programming (DP) problem.

The above DP model is for both the instantaneous and Coifman travel times due to the calculations of their link MSEs in Section 3.1. In fact, it is easy to see that the proposed DP model can be used for any other link travel time methods as long as the methods only depends on the starting and ending locations of the link.

4 Solution Algorithm and Complexity

In this section, we present a graph representation of the DP model. We start from the case that there is originally no sensor on the freeway and one needs to deploy K sensors. We then show in Section 4.3 how the graph can be revised to incorporate the case that there are $K' < K$

[3] Otherwise, suppose $(s'_k, y'_k), k_1 \leq k \leq k_2$ is the optimal solution instead for stage $k_1 \geq 1$ to $k_2 \leq K$. Then it is clear that $(\bar{s_k}, \bar{y_k})$, for $\bar{s_k} = s_k^*, 1 \leq k \leq k_1 - 1$ or $k_2 + 1 \leq k \leq K$, $\bar{s_k} = s'_k, k_1 \leq k \leq k_2$; $\bar{y_k} = y_k^*, 1 \leq k \leq k_1 - 1$ or $k_2 + 1 \leq k \leq K, \bar{y_k} = y'_k, k_1 \leq k \leq k_2$ will produce smaller objective than $(s_k^*, y_k^*), 1 \leq k \leq K$. This is a contradiction.

existing sensors, and how to place extra $K - K'$ sensors to achieve the best travel time estimation performances.

4.1 A Graph Representation

A graph representation of the DP model, is depicted in Figure 2. In the figure, stages are listed horizontally and sections are listed vertically. Since we deploy one sensor per stage, we associate each link with a stage as well. The state of a stage represents the starting section of the link associated with the stage. In this figure, all possible states of a stage are represented as *nodes*. The node number is the section number. For example, the node at Stage 2 and Section 2 represents that the starting location of Link 2 could be Section 2. As mentioned before, there is only one state in Stage 1 ($s_1 = 1$) and $(N - K + 1)$ states (from k to $N - K + k$) for Stage $k = 2, \ldots, K$. We further create a fake stage as Stage $K + 1$ that has only one fake state $N + 1$.

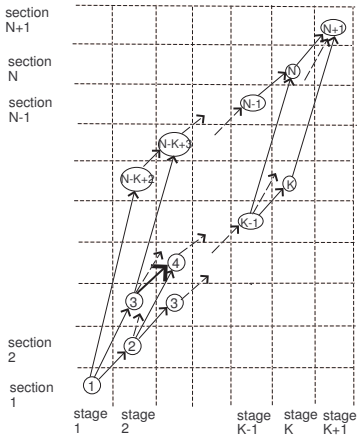


Figure 2: Graph Representation of DP Model

A connection may be created from a node in Stage k to another node in the immediate next state $k + 1$ if the latter node has a higher node number. This connection is denoted as an *arc* to distinguish with the roadway links associated with the sensors. Each arc actually represents a possible roadway link by defining the link's starting and ending sections. That is, an arc from Node s_k in Stage k to Node s_{k+1} in Stage $k + 1$ represents one possible configuration for Link k: it starts at Section s_k and ends at Section $s_{k+1} - 1$. Therefore, we must have $s_{k+1} > s_k$ in order to construct the arc. For example, the arc from Node 2 in Stage 2 to Node 4 in Stage 3 (marked in bold line) in Figure 2 means that one possible configuration for Link 2: it starts at Section 2 and ends at Section 3 (both are inclusive). Therefore, there should not be an arc from Node 4 in Stage 2 to Node 4 or lower in Stage 3. Further, there are no arcs between any two stages that are not adjacent to each other. We associate a cost with each arc in Figure 2. For the arc from Node s_k in Stage k to Node s_{k+1} in Stage $k+1$, the arc cost is $\hat{E}_k^i(s_k, s_{k+1} - 1)$ as computed in equation (7). In other words, the cost of an arc is the MSE of travel time estimation for its corresponding roadway link.

It is easy to check that the graph constructed in the above manner enumerates all possible states in each stage (1 to K) and all possible configurations (i.e., the starting and ending locations) of each link. It also incorporates all the constraints of the model shown in equations (12) - (15). More importantly, each path from Node 1 in Stage 1 to Node $N + 1$ in Stage $K + 1$ contains exactly K arcs, each of which represents a possible configuration of a particular roadway link (i.e., its starting

and ending sections). In other words, each path represents a potential sensor deployment scenario. Therefore the optimal sensor locations can be achieved by finding the minimum-cost path from Node 1 in Stage 1 to Node $N+1$ in Stage $K+1$. Since all arc costs are positive, the DP model proposed in this article can be solved by a shortest-path search algorithm.

Figure 3(a) depicts a small example to illustrate how the graph can be constructed. In this figure, we deploy 3 sensors on a segment with 6 sections, i.e. $K=3, N=6$. Then we have 4 stages and 7 sections. Stage 1 has one state at the first section, stage 2 has four states (sections 2, 3, 4, and 5), and stage 3 also has four states (sections 3, 4, 5, and 6). The fake stage 4 has only one state at the fake section 7. Arcs can only be added from nodes in stage 1 (or 2 or 3) to nodes in stage 2 (or 3 or 4) and from lower-numbered sections to higher-numbered sections. The cost of an arc from section s_k at stage k to section s_{k+1} at stage $k+1$ is $\hat{E}_k^i(s_k, s_{k+1}-1)$ as defined in (9). The nodes, arcs, and the costs associated with arcs complete the graph corresponding to the DP model. In this graph, any path from node 1 in stage 1 to node 7 in stage 4 contains three arcs. Each arc actually corresponds to a physical roadway link, and the path represents one possible link configuration (i.e. sensor deployment strategy). For example, we highlight in bold line the path $1->4->5->7$. The first arc starts at section 1 and ends at section 4, implying that the associated roadway link starts at section 1 and ends at section 3 (the next link starts at section 4). The second arc starts at section 4 and ends at section 5, meaning the second link contains only section 4. Similarly, the third link contains sections 5 and 6. It is easy to check that the graph enumerates all possible paths and the optimal strategy is represented as the shortest path from node 1 in stage 1 to node 7 in stage 4.

4.2 Complexity of the Algorithm

The complexity of the shortest path search algorithm depends on the structure and size of the graph in Figure 2. In particular, the following theorem provides its complexity.

Theorem 1 *For the DP model (16) - (18), the following two statements are true.*

(a) The graph constructed in Section 4.1 for the DP model is acyclic;

(b) The DP model can be solved in polynomial time. In particular, the complexity of solving the DP model is $O(K(N-K)^2)$ if $K \geq 2$.

Proof. For (a), notice from the way the graph is constructed in Section 4.1 that all arcs are from lower stages to higher stages and from lower sections to higher sections. Therefore, cycles must not exist in the graph, i.e., the graph is acyclic.

To prove (b), we have already shown that the DP model can be solved via a shortest path search from Node 1 in Stage 1 to Node $N+1$ in Stage $K+1$. According to (a), the graph is acyclic, implying that the complexity of the shortest-path search is linear in terms of the number of arcs in the graph [25]. The number of arcs in the graph in Figure 2 can be easily calculated: from Stage 1 to Stage 2 or from Stage K to Stage $K+1$, there are $N-K+1$ arcs. Between any other two stages, there are $\frac{(N-K+1)^2}{2}$ arcs. Therefore, the total number of arcs in the graph is: $2(N-K+1)+\frac{(K-2)(N-K+1)^2}{2}$. If $N \gg 1$ and $K \gg 1$, the complexity of the solution algorithm becomes $O(K(N-K)^2)$, which is polynomial. □

Theorem 1 is valid for the DP models for both the instantaneous and Coifman travel time methods. It is also easy to see that Corollary 1 below follows immediately Theorem 1.

Corollary 1 *If $N \gg K$, the complexity of the solution algorithm is $O(KN^2)$.* □

Theorem 1, especially Corollary 1, states that the complexity of solving the DP model proposed in this article depends linearly on the number of sensors to be deployed and quadratically on the number of sections. Further, it does not depend on the number of time intervals (i.e., the length of the study period). This implies that the proposed model can be efficiently solved, even for large-scale problems. For example, if we are to deploy 40 sensor to a 20-mile freeway segment, we have $K = 40$ and $N = 2112$ if the freeway is divided into 50ft small sections. According to Corollary 1, the complexity of solving the problem is $KN^2 = 40 \times 2112^2 \approx 1.8 \times 10^8$, which can be solved in seconds using standard computers. In addition, the DP model produces the exact solution for the optimal sensor location problem. Therefore, at least in theory, the proposed DP model and solution algorithm are more efficient than previous methods (e.g., the NN method in [19]).

4.3 Consideration of Existing Sensors

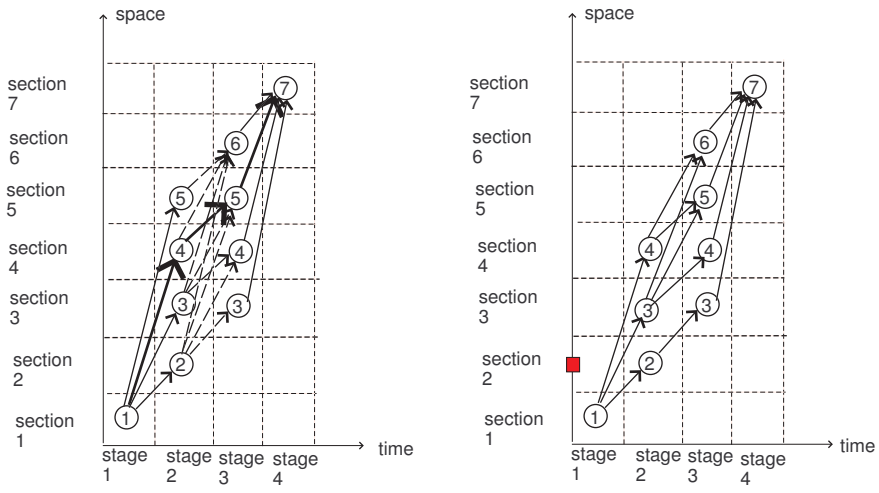
It may often be the case that one wishes to find the best way to add more sensors to a highway segment that already contains some existing sensors. In this case, we make a simple adjustment to the dynamic programming graph representation of the solution space. First, we match all existing sensors to the appropriate section they reside in. Then, every possible link (represented as an arc in the graph) that covers a section with an existing sensor in it but does not have the existing sensor at the center of the link is removed from consideration as a possible choice in the final solution. The reason for this is that we assume a sensor must be in the middle of its associated link.

As an example, imagine a highway section broken down into 6 sections. Then suppose that we already have a sensor in section 2. If this is the case, then we cannot consider links that cover section 2 but do not have section 2 as the middle of the link. This means that a link covering sections 1 through 4 would not be permissible in the solution (because that would imply a sensor in the boundary of sections 2 and 3 and not exactly on section 2), but a link covering sections 1 through 3 would be permissible. This can be further illustrated using Figure 3(b).

Therefore, to account for existing sensors, one can use a simple linear search on all of the links to identify which ones to remove and then uses the shortest path algorithm described in section (4.1) to compute the final solution. Therefore, the complexity of the algorithm remains, at worst, $O(KN^2)$.

5 Case Studies

In this section, we illustrate the proposed DP model and solution algorithm using two case studies. The first case study focuses on data obtained from micro-simulation, which provides an ideal situation since all individual trajectories are known. This allows us to investigate how the sampling rate (i.e. the percent of trajectories that are available to run the DP algorithm) will impact the sensor location quality. The second case study is based on real-world vehicle trajectory data from GPS-based cellular phones. They were obtained as part of the experiment to showcase the ability of using GPS cellular phones to collect and disseminate traveler information (**Alex, please add references to some mobile century papers here**).



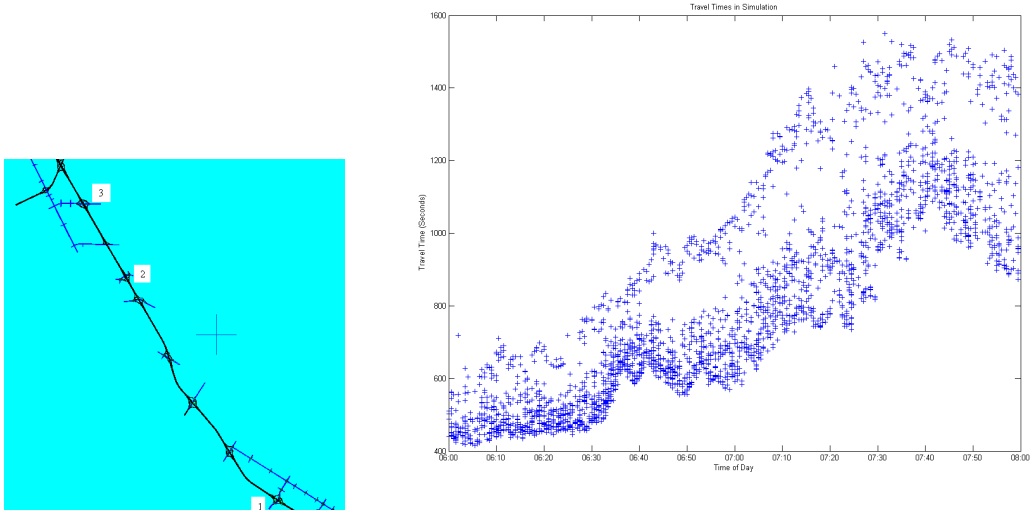
(a) An Illustrative Example (b) Consideration of Existing Sensors

Figure 3: Small Examples

5.1 Case Study I: Micro-Simulation Data Set

The micro-simulation data set is for a freeway segment that is roughly 8.7 miles in length from postmile (PM) 17.7 to 26.4. Figure 4(a) provides an overview of the network in Paramics, in which "1" and "3" indicates respectively the origin and destination of the route. We ran the simulation for 2 hours 30 minutes for morning peak hours from 5:30 am to 8:00 am. We then chose the last 2 hours as the study period, making the total number of 30-second time intervals equal to 240. We divide the freeway segment into 100-foot sections, resulting in $N = 459$. The "representative" vehicles are selected as those who traveled the entire segment and started their trips within the 2-hour study period. There are about 3,000 such vehicles, i.e., $M = 3000$. For all of those vehicles, the average travel time is 796 seconds and standard deviation is 227 seconds. Some basic characteristics of the travel times for this network is shown in Figure 4(b). We can immediately see that for a given time, the variation of travel times is large. In particular, if we use the average travel time at each 30-second interval as the base, the mean absolute variation of travel times is about 15%. This means that even if the instantaneous method or Coifman's method were able to perfectly predict the average travel time for each 30-second interval, we would still see about 15% error. Since each method is not perfect, we might consider the error above 15% the true error of the method.

We implemented the Dijkstra's shortest-path search algorithm [26] to solve the DP model. We varied the number of sensors from $K = 3$ to $K = 25$, or equivalently an average spacing from about 3 miles to 0.3 mile. We first depict in Figure 5 how the objective value computed by equation (2) decreases as the number of sensors increases from 3 to 25. The decrease is monotonic, but the marginal benefit decreases as well. We can also observe that the Coifman method usually has smaller objective value than that of the instantaneous method. As one example, Figure 6(a) depicts the obtained optimal sensor locations (marked using triangles on the y-axis in the figure) when $K = 6$ using the instantaneous travel time method. Note that in the figure, the speed contours of the segment are also displayed. Comparing with the freeway layout on the left side of the figure, we can first observe that this segment of freeway has two major bottlenecks. Both of them are due to merging located at about PM 26.0 and PM 23.5 respectively. In the latter half of



(a) Paramics Simulation Network (b) Travel Time Distribution

Figure 4: Simulation Network

the simulation, we can see that the first bottleneck propagates backward and combines with the second bottleneck. In this sense, we can treat them as a single congested area, spanning from PM 26 to PM 20. In addition, at about PM 18.5, there is a minor bottleneck for a short period of time (roughly from 7:20 am to 8:00 am).

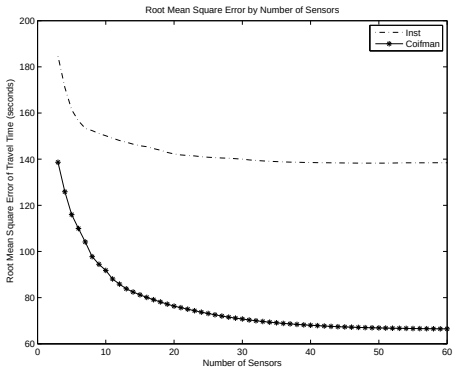
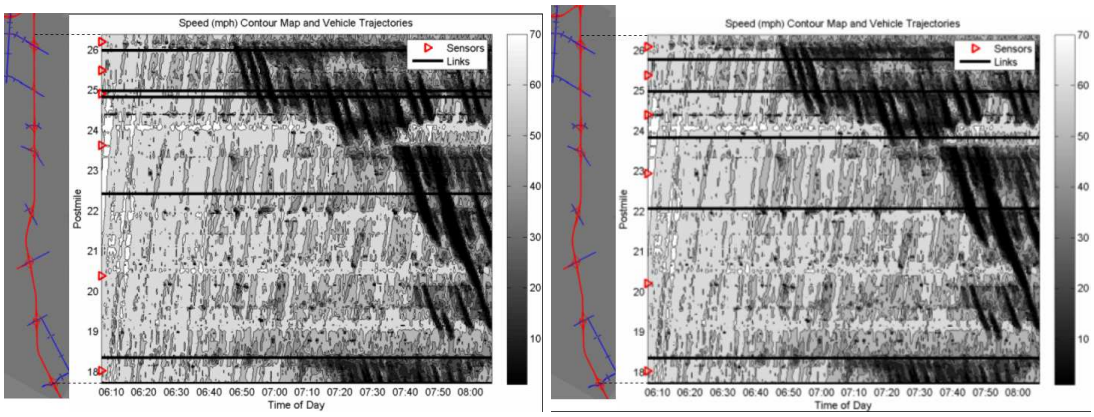


Figure 5: DP Objective Values vs. # of Sensors

The DP model using the instantaneous method puts four sensors at the major bottlenecks (PM 26.2, 25.5, 24.8, 23.5), one at the free flow regime (PM 20.4), and the last one at the minor bottleneck area (PM 18.0), which intuitively makes sense. We also ran the model using the Coifman method and the solution is shown in Figure 6(b). Cofiman's method generates similar results to the instantaneous method, i.e. four sensors in the congested area, one in the free flow area, and the last one in the minor bottleneck area. There are however some differences. In particular, the Coifman method tends to be able to distinguish the two major bottlenecks by putting three on the first bottleneck and one on the second bottleneck. This makes sense since the Cofiman's method constructs travel times by "walking through" both the spatial and temporal domains and thus is able to capture the dynamic evolution of bottlenecks. The instantaneous method on the other hand

only focuses on the snapshot of traffic conditions and is thus less sensitive to the actual shapes of bottlenecks.



(a) Instantaneous Method (b) Coifman Method

Figure 6: Optimal Sensor Locations for Simulation Data (6 Sensors)

To better illustrate how bottleneck areas impact the optimal sensor locations generated by the DP algorithm, we show in Figure 7 how the sensor locations change as we increase the number of sensors from 3 to 12, computed using the instantaneous travel time method. We can observe from this figure that when the number of sensors is small (e.g. $K = 3$), they will be first deployed to major bottlenecks (i.e. PM 25.5 and PM 23.7). For the free-flow area (e.g. at PM 20.1), only one sensor is needed. As more sensors are available, they will be deployed to bottleneck areas to capture the complicated traffic conditions in bottlenecks. Also, as the number of sensors increases, minor bottlenecks may also be captured and enhanced by extra sensors, while usually one sensor is sufficient for free flow areas. More importantly, as extra sensors are added in, the locations of previously deployed sensors in bottleneck areas remain almost unchanged. This is illustrated using the thin lines in the figure, which show that locations of newly deployed sensors just "branch out" from existing sensors in bottleneck areas. This implies that the DP algorithm has the ability to capture the most significant bottlenecks and if more sensors are available, the second significant bottlenecks will be captured and so on. The locations of sensors at free flow areas however may change since the speeds detected at free flow areas are not sensitive to the actual sensor locations. The evolution of optimal sensor locations via DP for the Coifman method is similar to that in Figure 7. The above discussions illustrate the close relation between the optimal sensor locations generated by the DP algorithm and the bottleneck areas of the network. They also show that the results from DP are stable and predictable, which is desirable in practice.

Notice that the DP objective function defined in (2) only looks at the summation of MSEs of individual links, instead of the MSE of travel times of the entire route. The latter can be defined as follows:

$$\bar{E} = \frac{\sum_{m=1}^{M} \left(\frac{\sum_{k=1}^{K} e_k^m}{\sum_{m=1}^{K} \tau_k^m} \right)^2}{M}. \tag{19}$$

Equation (19) defines the MSE in a relative sense since $\sum_{m=1}^{K} \tau_k^m$ is the actual travel time of the m-th vehicle and $\sum_{k=1}^{K} e_k^m$ is the estimation error for that vehicle. To show that the DP results are also (nearly) optimal for the objective in (19), we compare the objective values (computed by

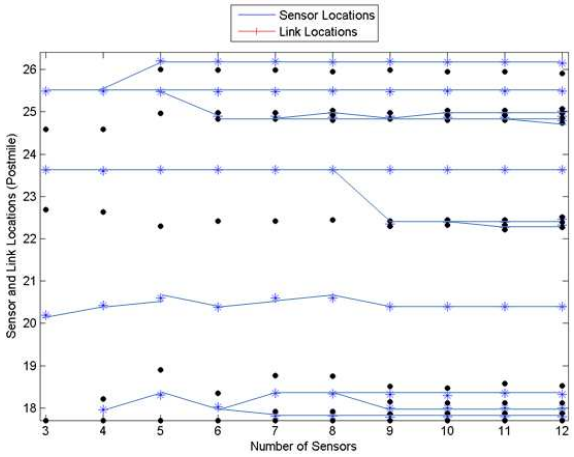


Figure 7: Evolution of Optimal Sensor Locations (Instantaneous Method)

by (19)) of the DP solution with those from 1,000 randomly generated sensor configurations for 2 - 25 sensors. The results are shown in Figure 8 for the instantaneous method. In this figure, the solid line represents the average objective values across all random configurations with the best and worst random configurations represented by the ends of the error bars. The line with rectangle signs represents objective values via evenly spaced configurations, while the line with asterisks represents the objective values from the DP solution. Clearly, the DP solution curve is very close to or lower than the smallest objective values by all random configurations. This indicates that the DP solution does generate near optimal solution even the objective function used in DP is (19). We can also observe that evenly spaced sensors cannot produce satisfactory travel time estimation results especially when the number of sensors is small. For example, when the number of sensors is $K = 3$, the objective value of the DP solution is 32%, while evenly spaced configuration produces 68% error. In addition, the performance of the evenly spaced configurations tend to vary significantly when the number of sensors varies. The performance of the DP solution however is very stable. These differences tend to reduce as the number of sensors increase. For example, $K = 25$, the objectives values for the DP solution and evenly spaced configuration become 28% and 37% respectively. This indicates that optimal sensor placement is more critical for limited number of sensors, while it is less critical if there are sufficient number of sensors to be deployed (in this case, evenly spacing the sensors may work pretty well).

In Section 2.2, we mentioned that the objective function defined in (2) can generate an optimal solution to the entire route that will likely be close to optimal even for its sub-routes. To illustrate this, we pick one sub-route as indicated in Figure 4(a) using "2" and "3". We then evaluate how the DP solution and the best random configuration (computed for the entire route) perform on this sub-route for each given number of sensors (2-25). This is displayed in Figure 9. In the figure, the solid line with asterisks represents the objective values calculated using (19) by evaluating the DP solution on this sub-route; the dashed line represents the objective values by evaluating the best random configuration (generated for the entire route) on the sub-route. The two curves show that the DP solution is consistently superior to the best random configuration on the sub-route. More importantly, the performance of the DP solutions is more stable across different numbers of sensors, while the random configurations tend to have varied performances depending on the actual number of sensors.

In reality, it is almost impossible to obtain trajectories of all vehicles traversing a given segment of freeway. Therefore, one crucial issue is to study how sampling rate impacts the results of the

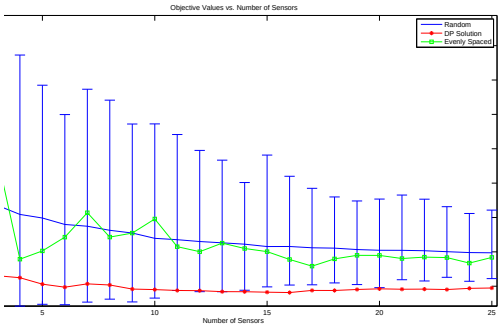


Figure 8: Objective Value vs. # of Sensors

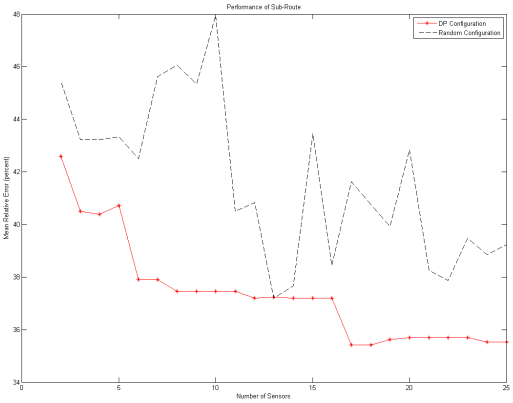


Figure 9: Performances on the Sub-Route

DP algorithm. Since DP solutions are closely related to bottleneck areas of the route as discussed above, we focus on the speed contour map for this purpose. In particular, we vary the sampling rate from 0.5% to 100%. For a given sampling rate α, we select each of the M total vehicles with probability α. The objective is the average difference between the speed contour map by all the vehicles and that by the sampled vehicles, which is defined as follows:

$$E_s = \sqrt{\frac{\sum_{n=1}^{N}\sum_{t=1}^{T}(v_{n,t} - \hat{v}_{n,t})^2}{NT}}. \tag{20}$$

Here E_s denotes the average absolute error of two speed contour maps, $v_{n,t}$ is the average speed for section n at time interval t computed using all vehicles, and $\hat{v}_{n,t}$ is the average speed for section n at time interval t computed using only sampled vehicles.

Figure 10 shows that the error decreases quickly from 0.5% to 25%, after which point the error is less than 1 mph. At 5%, the error is about 2 mph. Therefore, 5% to 10% seems to be a reasonable range in which one would expect the speed maps by the sampled vehicles to be very close to that from all of the vehicles. This is consistent with previous studies (e.g. 4% is concluded as sufficient for travel time estimation in [27]).

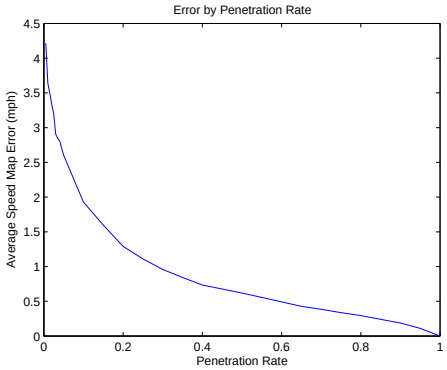


Figure 10: Sampling Rate

5.2 Case Study II: GPS-Equipped Cellular Phones Data

We further validate the DP model and algorithm using trajectories obtained from GPS-equipped cell phones. The data are from the Mobile Century demonstration, which deployed 100 cars equipped with GPS-enabled cell phones to loop over a 5.5 mile freeway segment of Interstate 880 in the San Francisco Bay Area [**Alex, please elaborate more and add appropriate references**]. The experiment was conducted from 9:00 am to 7:00 pm on February 8, 2008. Trajectories of all the 100 vehicles were collected via GPS over the entire experiment duration. Figure 11 illustrates the experiment site. Note that in the figure, "1" and "3" indicate respectively the origin and destination of the NB travel. The average travel time of the loop is about 20 minutes, implying that the 100 experiment cars represent about 300 vehicles/hours extra freeway traffic volume. Since the freeway has three through lanes in this area, the capacity of the freeway is roughly 6000 vehicles/hour. In other words, the resulting sampling rate of the obtained trajectories is about 5%. In this article, we focus on the northbound of the loop from 10:15 am to 1:45 pm.

Figure 12 (a) and (b) show the speed contour map for the freeway segment (NB) generated by loop detector data (from PeMS, pems.eecs.berkeley.edu) and GPS data respectively. We can observe that the GPS data can reproduce almost exactly the same speed contour as the detectors do. This verifies that 5% sampling rate is sufficient to capture speed contours or bottlenecks of the freeway, at least for the experiment. Notice that the bottleneck at 10:45 am was due to an accident on the freeway. The distribution of travel times collected by the experiment cars is shown in Figure 13. It can be shown that the variation of the travel times is about 4%, which is small compared with the simulation travel times in Section 5.1.

We first run the DP algorithm by varying the number of sensors from $K = 2$ to $K = 25$, or equivalently for an average spacing from about 3 miles to 0.2 mile. Figure 14 depicts how the objective value used in the DP model changes as the number of sensors increases. Similar to the results for simulation data, the objective value decreases as the number of sensors increases, and the Coifman's method always has smaller objective values. As one example, Figure 15 depicts the obtained optimal sensor locations when $K = 6$ using the instantaneous travel time method. Similar to the results in Section 5.1 for simulation data, the DP method puts most sensors (5) to the only bottleneck at the far north of the segment, while only one sensor is deployed to the free flow area (2 sensors when the number of sensors is 10 or 11). Furthermore, if we look at the evolution of optimal sensor locations as the number of sensors increases from 2 to 12, as shown in Figure 16, similar observations can also be obtained: most sensors are deployed to the major bottleneck area and only one sensor to the free flow region; as the number of sensors increase,

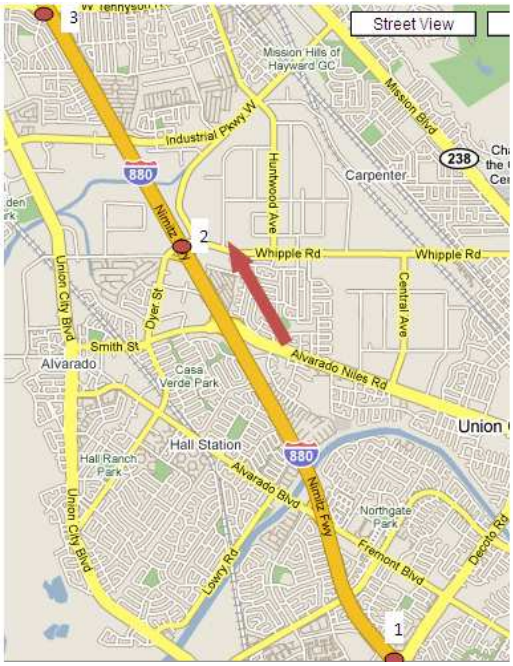


Figure 11: Experiment Site

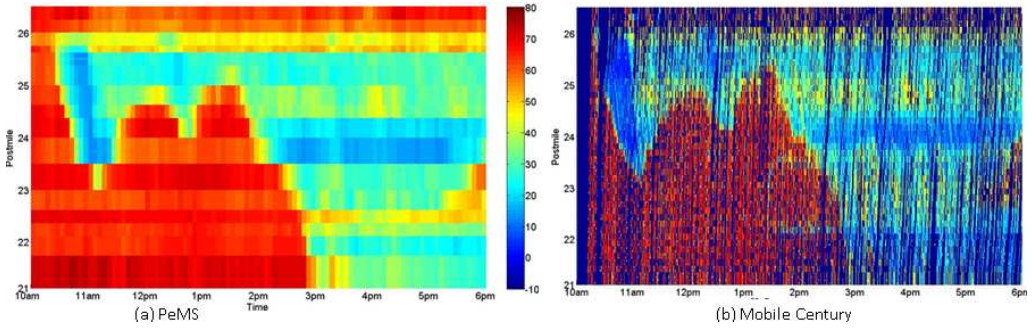


Figure 12: Speed Contour Maps

previously deployed sensors in the bottleneck area remain almost unchanged and new sensors just branch out from existing sensors. Same results can also be obtained for the Coifman method.

We compute the objective values as defined in equation (19) for the DP solution and 1,000 randomly generated sensor location configurations. Figure 17 depicts that the DP solution is near-optimal compared with the best random configuration for any given number of sensors. Again, the performance of evenly spaced configurations varies significantly and cannot compare with the DP solutions, whose performance is very stable. We then obtain the sensor locations for the entire route using the proposed DP model and 1000 random sensor configurations, and evaluate the objective value (19) on the sub-route defined in Figure 11 (i.e. from "2" to "3"). The results are shown in Figure 18. The best random configurations vary significantly in performance and are inferior to the DP solutions. This once again verifies that the DP solution also works well for sub-routes of the study freeway segment.

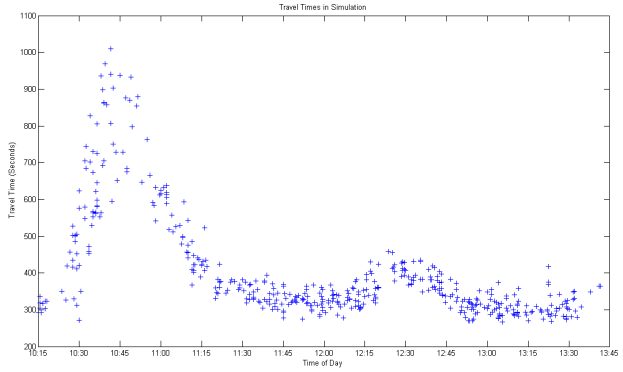


Figure 13: Travel Time Distribution of GPS Data

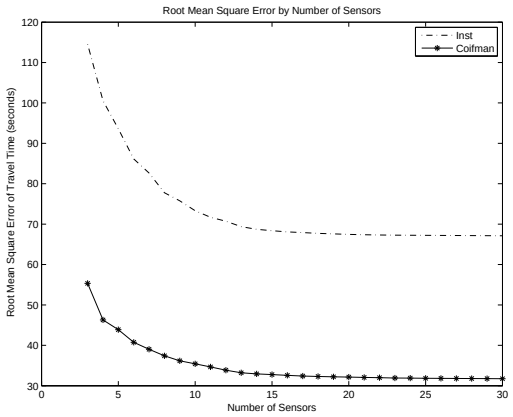


Figure 14: DP Objective Value vs. # of Sensors

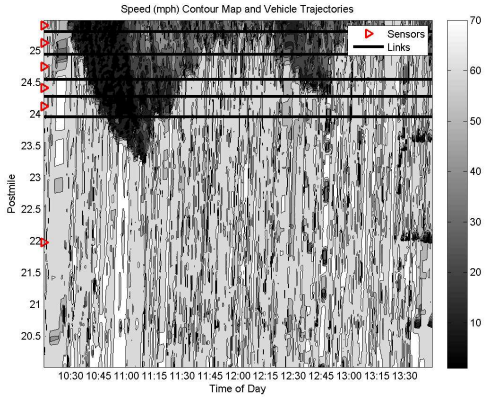


Figure 15: Optimal Locations for GPS Data for 6 Sensors

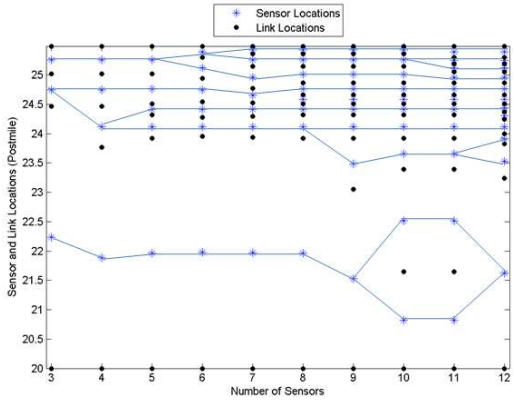


Figure 16: Evolution of Optimal Sensor Locations

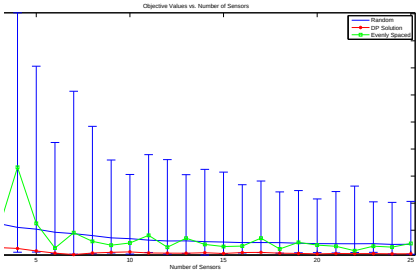


Figure 17: Objective Value vs. # of Sensors for GPS Data

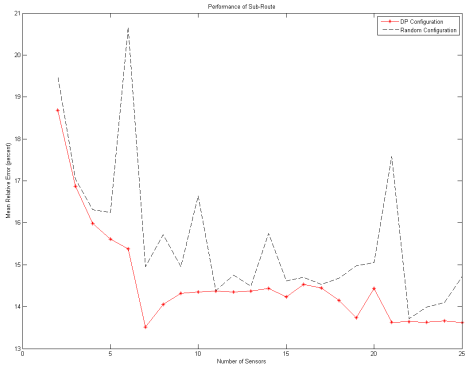


Figure 18: Performance of Sub-Routes for GPS Data

6 Conclusion

We studied in this article the optimal sensor placement problem for providing freeway travel times. The study is based on the assumption that vehicle trajectories are available and sensors can only provide aggregated speeds. We showed that based on those assumptions, determining optimal sensor locations can be modeled as a dynamic programming (DP) formulation and solved using shortest-path search in an acyclic graph. Therefore, the proposed model can be solved

in polynomial time and can be applied even for large-scale problems. We also showed how to incorporate existing sensors in the proposed DP framework. We then provided two case studies based on trajectory data from the Mobile Century demonstration and micro-simulation. The results show that 1) it is optimal to place many sensors in bottleneck areas and place just a few sensors in free flow areas; 2) the DP solution is more reliable and predictable than random configurations, which is more desirable in practice; and 3) there seems to be an optimal number of sensors that should be deployed, and beyond which deploying more sensors is not very beneficial.

The DP model and solution algorithm are the first step in determining optimal sensor placement to provide freeway travel times. There are several issues that remain unanswered. Below are some of them:

1. How sensitive is the model to different travel time estimation methods? We only illustrated the proposed model and algorithm using two travel time methods. However, since the DP model is constructed using link estimated and actual travel times only, it is our understanding that the DP model can be also applied to other travel time methods as long as the route travel time is calculated using the summation of link travel times. Nevertheless, how sensitive the resulting sensor locations are to different travel time methods merits further investigations.
2. How sensitive the model is to different sets of vehicle trajectories? In this article, we only utilized trajectories for one simulation run or one experiment. Therefore, day-to-day variations are not considered. How different sets of trajectories will impact the "optimal" sensor locations is an interesting research topic. This issue is under investigation now and results will be reported in subsequent papers.
3. How to account for sensor errors? In reality, almost all sensor data are subject to detection errors. How to consider the sensor errors when determining sensor locations is a practical yet challenging problem. In this regard, quantifying the errors of different types of sensors (such as loop detectors, speed radar sensors, etc.) seems necessary. The DP model proposed in this article may also be extended for this purpose. Research in this direction will be pursued in the future.

References

[1] H. Yang and J. Zhou. Optimal traffic ccounting locations for origin-destination matrix estimation. *Transportation Research B*, 32(1):109–126, 1998.

[2] L. Bianco, G. Confessore, and P. Reverberi. A network based model for traffic sensor location with implications on o-d matrix estimates. *Transportation Science*, 35(1):50–60, 2001.

[3] Z.C. Li, W.H.K. Lam, S.C. Wong, H.J. Huang, and D.L. Zhu. Reliability evaluation for stochastic and time-dependent networks with multiple parking facilities. *Networks and Spatial Economics, in press*, 2007.

[4] H. Al-Deek and Emam B. Emam. New methodology for estimating reliability in transportation networks with degraded link capacities. *Journal of Intelligent Transportation Systems*, 10(3):117–129, 2006.

[5] A. Chen, H. Yang, H.K. Lo, and W. Tang. A capacity related reliability for transportation networks. *Journal of Advanced Transportation*, 33:183–200, 1999.

[6] C.D.R. Lindveld, R. Thijs, P.H.L. Bovy, and N.J. Van der Zijpp. Evaluation of online travel time estimators and predictors. *Journal of Transportation Research Board*, 1719:45–53, 2000.

[7] Y. Iida, N. Uno, and T. Yamada. Experimental analysis approach to analyze dynamic route choice behavior of driver with travel time information. In *Proceedings of Vehicle Navigation and Information Systems*, pages 377–382, 1994.

[8] A. Khattak, A. Kanafani, and E.L. Colletter. Stated and reported route diversion behavior: implications of benefits of advanced traveler information system. *Transportation Research Record*, 1464:28–35, 1994.

[9] J. Rice and E.V. Zwet. A simple and effective method for predicting travel times on freeways. In *Proceedings of IEEE Intelligent Transportation Systems*, pages 227–232, 2001.

[10] B.Coifman. Estimating travel times and vehicles trajectories on freeways using dual loop detectors. *Transportation Research A*, 36(4):351–364, 2002.

[11] K. Ozbay, B. Bartin, and S. Chien. South Jersey real-time motorist information systems: Technology and practice. *Transportation Research Record*, 1886:68–75, 2004.

[12] I. Fujito, R. Margiotta, W. Huang, and W.A. Perez. The effect of sensor spacing on performance measure calculations. In *Proceedings of the 85th Annual Meeting of Transportation Research Board (CD-ROM)*, 2006.

[13] J. Kwon, B. McCullough, K. Petty, and P. Varaiya. Evaluation of PeMS to improve the congestion monitoring program. Technical report, Final Report for PATH TO 5319, 2006.

[14] X. Ban, Y. Li, A. Skabardonis, and J.D. Margulici. Performance evaluation of travel time methods for real time traffic applications. In *Proceedings of the 11th World Congress on Transport Research (CD-ROM)*, 2007.

[15] G. Thomas. The relationship between detector location and travel characteristics on arterial streets. *Institute of Transportation Engineers Journal*, 69(10):36–42, 1999.

[16] S. Oh, B. Ran, and K. Choi. Optimal detector location for estimating link travel time speed in urban arterial roads. In *Proceedings of the 82nd Annual Meetings of the Transportation Research Board (CD-ROM)*, 2003.

[17] S.M. Eisenman, X. Fei, X. Zhou, and H.S. Mahmassani. Number and location of sensors for real-time network traffic estimation and prediction: A sensitivity analysis. *Transportation Research Record*, 1981:253–259, 2006.

[18] H.D. Sherali, J. Desai, and H. Rakha. A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. *Transportation Research B*, 40:857–871, 2006.

[19] B. Bartin, K. Ozbay, and C. Iyigun. A clustering based methodology for determining the optimal roadway configuration of detectors for travel time estimation. *Transportation Research Record (in press)*, 2007.

[20] Z. Jia, C Chen, B Coifman, and P Varaiya. Pems algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. In *Proceeding of IEEE ITS Annual Meeting*, pages 536–541, 2001.

[21] Berkeley Transportation Systems (BTS). Pems user guide, version 5.2. 2004.

[22] X. Ban, L. Chu, and H. Benouar. Bottleneck identification and calibration for corridor management planning. *Transportation Research Record*, 1999:40–53, 2007.

[23] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya. Detecting errors and imputting missing data for single-loop surveillance systems. *Journal of Transportation Research Board*, 1981:160–167, 2003.

[24] R. Bellman and S. Dreyfus. *Applied Dynamic Programming*. Princeton University Press, 1962.

[25] D. P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, 1998.

[26] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[27] K.K. Sanwal and J. Walrand. Vehicle as probes. Technical Report UCB-ITS-PWP-95-11, California Path, 1995.

Optimal Sensor Locations for Freeway Bottleneck Identification

Henry X. Liu, PhD
Assistant Professor
Department of Civil Engineering
University of Minnesota
500 Pillsbury Drive S.E.
Minneapolis, MN 55455
Phone: (612) 625-6347
Email: henryliu@umn.edu

Adam Danczyk
Graduate Research Assistant
Department of Civil Engineering
University of Minnesota
500 Pillsbury Drive S.E.
Minneapolis, MN 55455
Phone: (612) 625-0249
Email: danc0010@umn.edu

Revised Manuscript to Computer-Aided Civil and Infrastructure Engineering

January 20, 2009

ABSTRACT

In the field of traffic operations, accurate performance measures are crucial for many of the Intelligent Transportation Systems applications. Achieving this accuracy and quality requires that network-based roadway sensors are allocated in locations beneficial to traffic operations. However, with the budgetary restrictions most transportation agencies face, these roadway sensors cannot be placed as thoroughly as obligatory for ideal accuracy, requiring these agencies to select a limited number of installments that produce the most optimal results. In this paper, a non-linear integer program is proposed to optimally allocate point sensors along a one-directional freeway corridor, given that any pair of adjacent sensors can produce a benefit for bottleneck identification. The objective of this model is to optimize the accuracy of bottleneck identification subject to resource and monetary constraints. This model is nonlinear and, due to a non-differentiable variable, Genetic Algorithm is applied to find a solution. We demonstrate that on a case study network with bottlenecks at unknown locations, the model successfully allocates sensors in a manner that optimizes bottleneck identification accuracy.

INTRODUCTION

With numerous Intelligent Transportation System (ITS) applications currently in operation, the need for high quality and reliable data is tremendous. It is known that data reliability and accuracy is dependent on the location allocation of roadway sensors to measure conditions and it has been shown that refined accuracies can be achieved by increasing the number of measuring devices per unit length (Kwon et al., 2006). In an ideal, monetary-free world, one would instrument a given roadway with a spatially-maximizing number of sensors to find the "ground truth" traffic state. Unfortunately, this ideal world is not reality, as budgetary constraints often forbid such lavish instrumentation. Instead, practitioners are left to seek out the most optimal placement of sensors given their constraints. Finding this optimal placement in a highly complex traffic environment is not often an easy achievement.

Bottlenecks are one of the most influential forces for the degradation of a transportation network and can often be the leading cause of delay on a transportation network, as is seen in Chen et al. (2003), where the bottlenecks in the study caused 64% of the observed delay. These bottlenecks occur at sites where flow is constricted, including at on-ramps, off-ramps, weaving areas, lane drops, curves, hills, and various other locations. Dealing with these bottlenecks requires transportation agencies to have an awareness of the recurrent frequency and location through interpretation of roadway sensor data. However, for these agencies to make an accurate assessment of each bottleneck, the roadway sensors need to be placed in a manner that produces useful measurements.

The purpose of this research is to develop a method for properly allocating roadway sensors around freeway bottlenecks to improve the accuracy of freeway performance measures. Since traffic conditions vary most in turbulent regions, especially those near bottlenecks, it is beneficial to focus resources to these regions as to better measure the drastic changes in traffic behavior. Through proper placement of roadway sensors, transportation agencies would receive more truthful data for performance monitoring, which in turn would improve traffic operation activities overall, such as ramp metering or traveler information. The proposed method would serve as a planning tool, primarily for infrastructure development projects, for prioritizing the assignment of additional roadway sensors to a corridor with an existing sensor configuration.

This research is also intended to fill in the gap among other studies for optimal sensor allocation with a limited budget. Optimal sensor spacing is a subject that has been investigated in the past, but not necessarily for bottleneck-specific purposes. Ozbay et al. (2004) studied the quality of travel time estimation when compared with sensor locations under recurrent and non-recurrent congestion. Kwon et al. (2006) created an empirical model that relates roadway-based sensor spacing to the accuracy of measuring traffic congestion by studying how overall accuracy falters as the distance between sensors increases. Ban et al. (2007) showed similar results, illustrating that increasing the sensor spacing causes higher travel time estimation errors and higher variations in travel time reliability. Bartin et al. (2007) also focused on roadway sensor spacing, finding that the marginal gain of travel time accuracy decreased as the number of road-based surveillance units increased.

A study by Fujito et al. (2006) determined that the actual location of sensors is more important for the estimation of congestion along a transportation corridor than uniform spacing. Empirical analysis showed that results varied accordingly to the positions in which sensors were removed. This strategic location concept can be seen in numerous transportation sensor location problems. Several studies develop models for determining instrumentation locations on transportation network that help uncover origin-destination (O-D) travel demands (Fei and Mahmassani, 2007, Fei et al., 2007, Bianco et al., 2001, Yang and Zhou, 1998, Yang et al., 2006). Others focus on sensor allocation on a transportation network for toll collection purposes (Zhang and Yang, 2004) and for traffic monitoring to reduce network risks (Gendreau et al., 2000), such as policing for drunk drivers.

Sherali et al. (2006) developed a model for optimally allocating Automatic Vehicle Identification (AVI) tag readers along a transportation corridor. Their research assumed that an environmental characteristic, called benefit, exists between any two sites and can be captured by allocated sensors at these sites. In this case, the environmental characteristic was deemed to be travel time variability. Their model has a quadratic objective function and a set of linear constraints, and is solved using the Reformulation-Linearization Technique. While this model is a significant contribution, its application is limited to AVI tag readers or other reidentification sensors, which have traditionally not been the most common sensors used by transportation practitioners. The more popular forms of sensors, which include inductance loop detectors and other point sensors that

collect aggregated traffic flow data, have not been considered in their study. In practice, performance measures are assessed between two neighboring point sensors only, whereas performance measures can be assessed between any reidentification sensors. Consequently, there is a need to produce a model that can optimize the allocation of such point sensors for performance measuring purposes while balancing resource and monetary constraints.

This paper will start by discussing the background literature surrounding bottleneck characteristics and detection. From there, it will develop an analytical model that can determine sensor placement around a bottleneck and test the model on a simple scenario. Lastly, it will test this model on a case-study network to see if it can find bottlenecks in a complex, dynamic environment.

BOTTLENECK CHARACTERISTICS AND DETECTION

A good quantity of literature is available to discuss algorithms to detect bottlenecks or freeway conditions near bottlenecks. An active bottleneck is defined by an upstream queue and unrestricted flows present on downstream sections (Daganzo, 1997). According to Zhang and Levinson (2004), three main traffic characteristics are present at bottlenecks: Flow Drop, Queue Discharge Flow, and the Pre-Queue Transition Period. Cassidy and Bertini (1999) observed vehicle discharge flows averaging 10% lower than flows measured prior to the queue's formation, which is significantly higher than previous studies (Agyemang-Duah and Hall, 1999; Banks, 1990; Banks, 1991; Newman, 1961; Persaud, 1986; Hall and Hall, 1991; Persaud and Hurdle, 1991). Queue discharge

flows seldom deviate from the mean rate by more than five percent and change gradually over time (Cassidy and Bertini, 1999; Bertini and Cassidy, 2002), although another study suggests queue discharge flows decrease as upstream queues become longer (Koshi, 1992). Pre-Queue Transition Periods per breakdown tend to range from 3 to 32 minutes (Cassidy and Bertini, 1999; Hall and Agyemang-Duah, 1991), where the breakdown probability is a function of mainline flows (Athol and Bullen, 1973; Persaud et al., 1998; Persaud, 2001).

Zhang and Levinson (2004) considered the issue of bottleneck formation in their study when the upstream detector station is deemed congested while the downstream loop detector station is deemed uncongested for more than five minutes. Occupancy over a data collection interval (30 seconds) was used to measure congestion, where a minimum reading over 25 percent was considered congested and a maximum reading of under 20 percent was considered uncongested. Chen et al. (2003) create an algorithm that declares the presence of an active bottleneck if upstream speeds are greater than downstream speeds by 20 MPH (miles per hour) and those detectors are less than two miles apart. The algorithm also stipulates that upstream speeds need to be less than 40 MPH for an active bottleneck.

From the literature, it is clear that bottlenecks have some definitive characteristics. Upstream of the bottleneck, vehicle densities are higher while velocities are lower. Downstream of the bottleneck, vehicle densities are lower while velocities are higher. This is important to note as it can illustrate the impact a bottleneck has on a corridor,

based solely on the variation of conditions between locations. This type of variation would assist in determining where sensors should be allocated in order to optimize benefit. Further discussion in relating these varied conditions to optimized benefit can be found in the following sections.

FORMULATION OF OPTIMAL SENSOR PLACEMENT PROBLEM

A freeway corridor will be considered, as shown in Figure 1. This freeway receives an entry flow rate at its upstream point (q_0) during the study period *T*. Along the route, entrance ramps add additional flow (q_1, q_3, q_5) while exit ramps reduce flow (q_2, q_4, q_6). These input data is assumed to be known *a prior*. If the proposed method is to be applied for a freeway corridor with empty sensors, then freeway design volumes are required. Otherwise for a freeway with existing sensors, to prioritize the allocation of additional sensors, traffic volume data from existing sensors can be applied. Now, the question is, given the freeway geometry and entry volumes, what would be the optimal allocation of roadway sensors for the purpose of bottleneck identification?

To solve this problem, the freeway is divided into N cells. These cells do not necessarily have to be of equal length, but for purposes of simplicity they will be of equal length in this work. Each cell i is designated as a potential site to place a roadway sensor. Generally speaking, this roadway sensor would be located in the middle of the cell. We assume that there exists a benefit value in terms of identifying and measuring bottlenecks,

by locating a pair sensors at cell i and cell j. This benefit factor b_{ij} will be discussed in the next section.

An analytical model can then be formulated for the *Sensor Placement* (SP) problem as follows.

$$Maximize: \sum_{i \in N} \sum_{i \in N} b_{ij} x_i x_j \quad (1a)$$

$$Subject\ To: \sum_{j \in N} x_j \leq R \quad (1b)$$

$$\sum_{j \in N} C_j x_j \leq B \quad (1c)$$

$$x \quad Binary \quad (1d)$$

The objective function (1a) seeks to maximize total benefit based on the allocation of sensor resources to given cells. Constraint (1b) limits the number of sensors from exceeding a value, R, which represents the maximum number of sensors that can be placed on a corridor. Constraint (1c) limits the total cost of installing sensors at any site j with a unit cost, C_j, from exceeding the budget, B. Constraint (1d) sets x_j as a binary variable, where 1 represents the placement of a sensor at cell j and 0 otherwise. For this work, sensor measurement error is not considered. It will be assumed that all sensors measure accurately and are fully operational.

The objective function and constraints are both intuitive and similar to Sherali et al. (2006). The key to solving this problem is finding a way to identify the benefit factor.

Benefit Factor (b_{ij})

The benefit factor (b_{ij}) is a value intended to represent benefits gained by allocating roadway sensors at cell i and cell j, generally in terms of ability to accurately and completely capture the given measure of performance. The benefit factor only applies to a segment that lacks any other instrumentation—that is, a third sensor cannot be placed at cell k, where cell k is between cell i and cell j. In such a case, the segment between cell i and cell j would be divided into two segments and two benefit values would be assigned. Therefore the benefit will be assessed for two "adjacent sensors only" to avoid "double counting". In other words, the benefit of adding a new sensor will be captured by two benefit factors, one pairing with adjacent upstream sensor and the other with adjacent downstream sensor. The benefit will be doubly counted if a positive benefit value is generated between two non-adjacent sensors. It should be noted that such a benefit factor is inherently different with that of Sherali et al. (2006), in that a benefit value can be assessed between any pair of "re-identification" sensors.

To assess an appropriate value for a benefit factor, it must first be considered what type of measurement would be beneficial to identify a bottleneck. The traffic behavior at a bottleneck is for vehicles upstream of the bottleneck to be moving much slower than vehicles downstream of the bottleneck, as reflected in Chen et al. (2003), where the algorithm to detect bottlenecks is based on speed variations. Therefore, such positive speed gradients between cell i and cell j would be beneficial to capture, as they would be reflective of bottleneck characteristics. However, simple speed variation is not sufficient enough, as conceptually it is more beneficial to have sensors closer to bottlenecks rather

than far to minimize the time of detection. This can be seen in an idealistic situation around a bottleneck, where velocities at all upstream locations would become the same, as would velocities at all downstream locations. In such a case, if only speed gradients are used in b_{ij}, then it would be equally beneficial to allocate sensors near a bottleneck or far from a bottleneck, which is conceptually incorrect. Therefore, to make sites closest to bottlenecks more appealing, the stated benefit factor must incorporate a distance element to penalize increasing distances. Since the corridor is one-directional, meaning traffic flows only in one direction, the distance will always be non-negative.

With these two critical elements in mind, the benefit factor can be formulated as follows.

$$b_{ij} = \begin{cases} 0 & \sum_{k=i}^{j} x_k > 2, \quad S_j < S_i \\ \max\left\{0, \quad \dfrac{\sum_{t=0}^{T}(V_j^t - V_i^t)}{(S_j - S_i)}\right\} & \textit{otherwise} \end{cases} \qquad (2)$$

The benefit factor is the non-negative speed gradient between cell i and cell j, weighted by their distance. V_j^t is a time-dependent velocity measured at cell j for time t. The benefit factor takes the sum of the differences of time-dependent velocities over the study period, T, and divides by the distance between cells. As mentioned earlier, the benefit factor only considers segments eligible for benefit if there are no other sensors present between cell i and cell j. If a sensor is present between i and j, the number of sensors between i and j would be greater than two, thus assigning a benefit of zero. This

maintains the “adjacent sensors only” exclusivity mentioned earlier. Additionally, since the corridor is one-directional, benefit may only be captured when the stationing, or physical location, of site i, S_i, precedes the stationing of site j, S_j, along this corridor.

For purposes of the demonstrations to follow, measuring the velocities in eligible cells would be done through the cell transmission model proposed by Daganzo (1994) and Daganzo (1995), since a similar cell-based concept is used to define locations for sensor placement. The cell transmission model divides a given corridor into cells with cell lengths determined by free-flow travel distances over a designated time increment. When entry traffic flows are given, the cell transmission model determines time-dependent flows and densities based on flows of vehicles entering and exiting each cell. This model is interactive, meaning conditions of adjacent cells pose consequences to conditions of the cell in question, thus following the behavior of a macroscopic traffic flow. Use of this model will capture the recurrent congestion conditions formed by bottlenecks, which is the primary focus of this work.

The cell transmission model determines the time-dependent number of vehicles, n, in any given cell i through the following formula.

$$n_i(t+1) = n_i(t) + y_i(t) - y_{i+1}(t) \qquad (3)$$

In the formula, time-dependent flows entering cell i at time t are identified as $y_i(t)$. These flows are based off a trapezoidal-shape flow-density model. This trapezoidal model is constrained by capacity, or Q_{Max}. as shown in Figure 2.

With this model, the volumetric flow can be calculated for any time t, as described in the following formula.

$$y_i(t) = \min\{n_{i-1}(t), Q_i(t), N_i(t) - n_i(t)\} \quad (4)$$

Equation (4) states that volumetric flow moving between cells is constrained by available vehicles in the upstream cell, flow capacity between those cells, or downstream cell spare capacity, respectively. $Q_i(t)$ represents the capacity flow into cell i for time interval t. $N_i(t)$ is the maximum number of vehicles allowed into cell i at time t.

Time-dependent velocities in cell i can be calculated by dividing densities by flows in the same interval and cell. These velocities will then be summed over the entire evaluation period to determine the benefit factor.

Solution Algorithm

The SP problem is a linearly constrained mixed-integer zero-one quadratic programming problem. However, the complexity of the non-differentiable benefit factor makes it difficult to solve with traditional nonlinear solvers. Consequently, a heuristic is necessary to find a solution. In this paper, genetic algorithm (GA) was chosen as the solution

heuristic, as the chromosome structure is suitably applied to zero-one problem at hand. Genetic algorithms operate as a simple system that imitates the natural selection process of genetic systems, where the fittest traits survive. While they are not guaranteed to find the optimal solution without fully enumerating all possible solutions, they are often successful in acquiring a solution with a high fitness in a reasonable time frame and converge towards a global optimum at an asymptotic rate. Holland (1975) proved the convergence of genetic algorithms. Goldberg (1989) discussed the operation of genetic algorithms in much more detail. GA has been widely used in civil engineering research literatures, goes as far back as 1993 (Adeli and Cheng, 1993).

To represent sensor allocation, a binary vector is translated into a chromosome, similar to the work done by Arafeh and Rakha (2005). Each gene represents an eligible site for roadway sensor allocation. Genes are coded as either 0 or 1, where 0 represents no sensor allocation and 1 represents sensor allocation. For this procedure, a population of thirty chromosomes was used. At the beginning, the number of budgeted roadway sensors was selected and an initial guess was made. Qualified chromosomes in the population are ones with very specific gene configurations. That is, the only chromosomes eligible to be included in the population are those with genes that summed up to the number of allowable sensors. In a scenario where eight eligible roadway sensor sites are present and the budgeted number of roadway sensors is four, then:

- A sample of acceptable chromosomes would include: 01010101, 11110000, 00001111.

- A sample of rejected chromosomes would include: 00000001, 00010101, 11111110.

With the population of acceptable chromosomes, a single-point crossover was conducted, following the example in Table 1. The crossover point was selected at random. To ensure that no roadway sensors were added or subtracted to the system, the chromosomes eligible for crossover must have the same number of genes with value “1” before and after the crossover point. If an original chromosome is 00110-010 (the dash representing the crossover point), then:

- A sample of eligible matches would include: 11000-001, 00011-100, 10001-001.
- A sample of ineligible matches would include: 01000-100, 00011-000.

Once two eligible chromosomes are identified, they are selected as the candidates for crossover. Using the same crossover point for determining eligibility, a new generation of chromosomes is developed. This is known as the reproduction process. The fitness of each offspring can be calculated by summing the total benefit between allocated sensors. A fixed number of chromosomes with the highest fitness values are selected to advance to the next generation. The chromosomes with lower fitness values are discarded. This is known as elitism. A technique known as mutation was also used in this research, where genes were randomly changed on an infrequent basis. Mutation is used to deal with local optima.

In this paper we ran through 1000 generations and considered the end result to be the best known. This run was conducted 10 times, each time using a different initial population to better search out the global optimum.

CONCEPTUAL EVALUATION WITH A KNOWN BOTTLENECK

To demonstrate the functionality of the SP program, a test will be performed on a simple situation where the presence of a freeway bottleneck is disrupting the traffic stream. In this case, the location of the bottleneck is known. The purpose of performing this simple test is only to illustrate that the SP program produces sensible results. A test on a complex traffic environment will come in the following section.

For this scenario, a simple, hypothetical pipeline freeway with a length of 25,000 feet and a free-flow speed of 68 MPH will be simulated. The time interval used for the cell transmission model will be 5 seconds, as to create cell lengths of 500 feet. The three-lane freeway will receive a capacity flow of 6,000 vehicles per hour (2,000 vehicles per hour per lane) for an hour. At 12,750 feet from the beginning of the segment, a short lane drop is present, generating a bottleneck when traffic volumes are sufficiently large. This single lane drop will reduce capacity at the site to 4,000 vehicles per hour. This lane drop only extends a short distance before returning to the earlier size and capacity.

With a length of about 25,000 feet and a cell length of 500 feet, 51 cells can be found on this corridor. Each of the 51 cells is considered an eligible site for a roadway sensor

allocation. As a result, the maximum number of sensors that can be placed, R, will equal 51. Average speeds in these cells will be measured for each time interval and then summed up over an entire hour. The cost, C_j, of installing a sensor at any cell j will be 1. For purposes of this study, we assume that traditional inductance loop detectors are used as traffic sensors.

Traffic conditions, starting with an empty network, were simulated for fifteen minutes before the one hour of data collection began. The intent of simulating conditions before data collection was to allow the freeway network to fill with vehicles and better exhibit the stable traffic conditions.

Four trials were conducted, each trial using a different number of detectors that can be placed, to observe how the SP program places these detectors. The results of these trials are illustrated in Figure 3, which shows where the SP program would assign loop detectors for a given number of budgeted loop detectors. Conceptually, for one bottleneck, it would be expected that one loop detector would be placed upstream and the other downstream, as to capture differing conditions and identify the bottleneck. This becomes true when two detectors are allowed. With this configuration, the bottleneck can be detected and measured. Since the freeway, simulated for fifteen minutes prior to data collection, is filled close to the long-term equilibrium state, the only variation in traffic speeds occur across the bottleneck, making the only benefit gained on the corridor be from the detectors placed immediately upstream and downstream of the site. This can

be seen in cases where more than two detectors are allowed, as they are allocated randomly without offering any additional benefit.

What would happen if the scenario became more complicated? If two bottlenecks were present and one was significantly more influential to the traffic stream than the other, how would the SP program assign loop detectors if constrained by budget? To demonstrate how the program behaves, the pipeline freeway scenario will be expanded upon. Keeping the original lane-drop bottleneck, an additional bottleneck will be added further downstream at 19,250 feet from the segment's beginning. This bottleneck will be a reduction of two lanes, reducing capacity through that bottleneck to 2,000 vehicles per hour. Corridor length and volumetric flow will be kept the same. Traffic conditions will be simulated for fifteen minutes prior to the start of data collection, as to fill the network in a similar fashion to the previous scenario.

How will the SP program react to the presence of a stronger bottleneck? Figure 4 illustrates the placement of detectors under this new scenario. The SP program behaves as expected, allocating loop detectors to the site of the stronger bottleneck. With two budgeted loop detectors, the SP program successfully places them immediately upstream and downstream of the double lane-drop bottleneck. When three loop detectors are made available, the SP program responds by placing an arbitrary detector because no other place in the network offers a marginal benefit increase. Upon receiving four detectors for use, the SP program places a detector pair at each bottleneck, thus illustrating the program's ability to deal with the strongest bottlenecks first. It is important to note that

the overall benefit has a relative increase of 18% as the fourth detector is added and the second bottleneck becomes detectable. A fifth detector is arbitrarily placed, as it also offers no additional benefit.

The marginal benefit increase for the bottleneck at 12,750 from the segment beginning is smaller in the two-bottleneck scenario than the single-bottleneck scenario. This difference is because the more impactful bottleneck overtakes the less impactful one during the hour of observation, therefore the less impactful bottleneck is active for less time and, thus, the benefit at allocating at that site is decreased.

Varied traffic volumes and densities do not hinder the ability of the SP program to properly allocate loop detectors. As long as the bottlenecks impact the traffic corridor and cause a disparity in speeds, the SP program will detect them and allocate loop detectors based on influence.

This section has shown the SP program's ability to allocate loop detectors for idealistic scenarios where a bottleneck has formed. In the next section, the SP program will be placed in a situation with complex traffic environments and multiple overlapping bottlenecks.

A CASE STUDY

How would the SP program allocate sensors in a more complex environment where locations of bottlenecks are unknown? To answer this question, the SP program will be

tested on a simulated freeway. This network represents westbound Interstate 94 (I-94) in the Minneapolis-St. Paul metropolitan area, running from Snelling Avenue to Interstate 394 (I-394) and totaling 7.2 miles. Along this path are 10 exit ramps, 7 entrance ramps, the presence of weaving areas, and freeway widths varying between 3 and 4 lanes, as shown in Figure 5. Current loop detectors on I-94, described as dark vertical lines in Figure 5, are placed approximately every half mile. This network was simulated using the cell transmission model for actual conditions from 2:00 P.M. to 8:00 P.M. local time on Wednesday, June 13, 2007. This time frame includes the PM Peak rush hour. Network volumes were determined by loop detector measurements on entrance ramps, exit ramps, and the starting point for the corridor of interest.

The time interval used for the cell transmission model is 6 seconds, as to create a cell length of 441 feet at a free-flow speed of 60 MPH. The total number of cells is 87, allowing the maximum number of loop detectors eligible for placement at this site, R, to equal 87. The unit cost, C_j, for placing a loop detector in any cell j will be 1. Several trials were conducted, where each trial had a different number of loop detectors that can be placed, to observe how the SP program places these detectors with different budgetary constraints. The first comparison, however, looked at where existing loop detectors, from Figure 5, can be better allocating, according to the SP program.

As is, the existing loop detector configuration on I-94 produces an overall benefit of 86.8, using the volumetric flows for the observed time. To seek out an optimal result, the SP program is used to relocate the 16 detectors for bottleneck identification and

measurement purposes. The results for the I-94 corridor are shown in Figure 6 and are what was expected. The majority of bottlenecks occur near downtown Minneapolis, where heavy traffic from the Interstate 35W (I-35W) interchange enters I-94. The other clustering of bottlenecks occurs between Snelling Avenue and T.H. 280, where traffic volumes in the real world are turbulent due to the busy entrance and exit ramps in the area. This new configuration generates an overall benefit of 1085.5, a significant increase over the previous configuration. In this case, sensors are paired around sites that create a high value of benefit, mostly on-ramps and off-ramps. This pairing is not necessarily typical, as different, complex sequencing of sensors has been observed in more complicated or turbulent environments.

Figure 7 illustrates the overall benefit results for the I-94 for an assortment of budgeted detectors, assuming no other detectors are present on the corridor. One series represents the benefit received using the configuration recommended by the benefit-maximizing SP program. The other series represents the benefit received if varied numbers of detectors were uniformly distributed across the corridor without any preference or bias. One point is included to represent the benefit gained by the current existing detector configuration.

As seen in Figure 7, the SP program performs better, in terms of benefit, than the uniform-spacing or the current configuration. For the SP program, the marginal benefit decreases to zero as the number of detectors increases and the bottlenecks are covered. This occurs at approximately 30 detectors, where any additional detectors bring miniscule incremental benefit. It comes as no surprise that the SP program allocated detectors to

merging and diverging areas, locations which have a high probability of creating a bottleneck.

For the uniform-spacing configuration, marginal benefit tends to increase as more detectors are placed on the corridor. Without strategic placement, these detectors will only measure bottleneck conditions if their configuration happens to place them near a bottleneck site. As the number of detectors increases, the likelihood of a bottleneck location being found increases, hence explaining the rising marginal benefit with each additional detector. The current detector configuration on I-94 happens to fall along this line, suggesting the current configuration bears similarities to being evenly spaced.

It is interesting to note the differences between the current configuration of loop detectors on I-94 and the configuration suggested by the SP program. A heavy clustering can still be observed near downtown Minneapolis and the I-35W interchange in the real world. However, it is clear that the real world spacing follows guidelines for other traffic operations, such as ramp metering or travel time estimation, as these loop detectors do not gravitate to where the bottleneck is estimated to exist.

The examples illustrated thus far have been on corridors without the presence of existing detectors. In reality, it is more likely that sensors would be added to a corridor that already contains an existing sensor configuration, as to further enhance the accuracy of performance measures. The SP program is applicable to such a corridor, serving as a tool to recommend where supplemental detectors should be located for bottleneck detection purposes based on the existing configuration.

The only variation resulting from applying this model to an existing network is that new constraints are required on the model, setting the binary variable x_j to a value of 1 in any cell j that has an existing detector. After adjusting the monetary budget to account for these existing detectors, it is clear that the model is still fully functional.

As stated earlier, 16 loop detectors exist on the I-94 corridor between Snelling Avenue and I-394, the segment of analysis in this paper. The existing benefit, as reported by the SP program, was 86.8. If the transportation agencies responsible for this corridor were to add additional detectors, what would be the resulting increases in benefit?

As shown in Figure 8, the benefit increases sharply when the first additional detector is added to the existing 16-detector corridor, more than doubling the total benefit. Then, as was the case in Figure 7, the marginal increases in total benefit begin to decrease in magnitude as the strongest bottlenecks receive sensor coverage first and then the weakest.

This section has shown the SP program's ability to allocate sensors on a corridor. As described, deploying additional sensors while using the SP program provides a substantial increase in the ability to identify and measure bottlenecks. The following section illustrates how well engineers would be able to detect bottlenecks given this new sensor configuration.

PERFORMANCE MEASURES

The success of the SP program finding an optimal allocation is dependent on how well that allocation truly identifies and measures bottlenecks. To illustrate this performance, a corridor performance measure is defined as Bottleneck Activity Time (BAT). BAT is a measure, in hours, of how much bottleneck activity is present along a corridor. For example, if one bottleneck appears for thirty minutes on a corridor receiving an hour of observation, then BAT would be 0.5 hours. Similarly, if three bottlenecks appear in that hour—each bottleneck having a lifespan of thirty minutes—the BAT would be 1.5 hours.

To determine BAT, three conditions are set as guidelines. First, a bottleneck is considered active when the speed measured at the downstream detector is at least 10 MPH greater than the speed measured at the upstream detector site. Second, this 10-MPH variation must have occurred consistently for at least five minutes before an inactive bottleneck can be considered active. Third, this variation must show a consistent measure of being less than 10 MPH for at least five minutes before an active bottleneck can be considered inactive again.

Figure 9 compares the absolute relative errors between the SP configurations and the uniform-spacing configuration in terms of BAT when compared with the ground truth state. The ground truth state, in this case, was found by deploying sensors to all sites and measuring the state. As seen, the SP configuration produces highly accurate estimates for BAT, far exceeding the accuracy of the uniform-spacing configuration. This makes sense conceptually, as the SP's objective of locating sensors closely to bottlenecks would allow these disruptions to be discovered in a much quicker time interval. From this data, it can be gathered that only three significant bottlenecks (with 10+ MPH variations on regular basis) exist on this corridor, requiring six strategically-placed detectors to make an assessment.

Figure 10 illustrates the absolute relative errors for the SP configuration when the model is applied to an existing configuration. As seen, only five additional, strategically-placed detectors are required to allow the existing 16-detector corridor to detect the significant bottlenecks (with 10+ MPH variations on a regular basis) present. This goes to show that the model can effectively find solutions for identifying bottlenecks with an optimal number of detectors, regardless if the corridor in question has existing sensor coverage or no coverage whatsoever.

CONCLUSIONS AND FURTHER RESEARCH

This paper has addressed the issue of allocating point sensors along a one-dimensional corridor to optimize the accuracy of bottleneck detection, given that point sensors follow a set of restrictive rules in terms of how bottlenecks can be detected. An optimization

model was proposed for optimally allocating sensors along this corridor given a set of known background conditions. This model was tested on a simple scenario with known bottlenecks to demonstrate its ability to allocate sensors correctly and prioritize allocation based on which bottlenecks are more detrimental to freeway efficiency. It was further tested on a simulated network with conditions similar to those found in the real world during an afternoon rush-hour period. This experimental test revealed the model's ability to locate bottlenecks in complex, changing traffic conditions and allocate sensors in appropriate locations for bottleneck identification. It also revealed the data accuracy improvement that practitioners would receive from such an optimized configuration when compared with a simple uniformly-spaced configuration. This problem is not a trivial one, as the sequences of sensors required in complex infrastructure and traffic environments may not be easy to determine through simple human intuition.

The work done in this paper has widespread applications in a variety of fields. It is not solely limited to bottleneck identification, as redefining the benefit factor would allow other performance measures to be optimized. Furthermore, the proposed model can be applied to both new and existing infrastructure projects. On an operations level, it would be assistive to the traffic operations and infrastructure community by answering the following questions:

- Given an existing network with a certain sensor configuration, where should new sensors be allocated when made available as to better measure the traffic state?

- When dealing with an existing network with ailing sensors present and a limited budget available for replacement, which sensors should be prioritized for replacement?
- Given a new, sensor-free road, such as a roadway arterial transformed into a freeway, how should sensors be spaced to identify existing bottlenecks?

In practice, allocation of roadway sensors to bottleneck sites may impede the ability of transportation agencies to conduct other operations activities, such as ramp metering or travel time estimation. As different traffic operations applications may have different requirements for sensor deployment, some of which may potentially be in conflict with each other, a multi-objective sensor location problem should be formulated to balance the need for different applications. This is left for future research.

ACKNOWLEDGEMENTS

Financial support for conducting this research was provided by California Department of Transportation (Caltrans) and the University of California, Berkeley. Sincere thanks go to Mr. JD Margulici and Dr. Jeff Ban of UC-Berkeley for their contributions and collaboration in this project.

REFERENCES

Adeli, H. and Cheng, N.-T., (1993) Integrated Genetic Algorithm for Optimization of Space Structures, *Journal of Aerospace Engineering*, ASCE, Vol. 6, No. 4, pp. 315-328.

Agyemang-Duah, K. and Hall, F.L. (1999), Some Issues regarding the Numerical Value of Capacity, Proceeds of the International Symposium of Highway Capacity, A.A. Balkema Press, Germany, pp. 1-15.

Arafeh, M. and Rakha, H. (2005), Genetic Algorithm Approach for Locating Automatic Vehicle Identification Readers, Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria.

Athol, P., and Bullen A. (1973), Multiple ramp control for a freeway bottleneck, Highway Research Record No. 456, pp. 50-54.

Ban, J., Li, Y., Skabardonis, A., Margulici, J.D (2007), Performance Evaluation of Travel Time Methods for Real Time Traffic Applications. In *Proceedings of the 11th World Congress on Transport Research*.

Banks, J.H. (1990), Two-capacity Phenomenon at Freeway Bottlenecks: A Basis for Ramp Metering?, Transportation Research Record 1320, pp. 83-90.

Banks, J.H. (1991), Flow Processes at a Freeway Bottleneck, Transportation Research Record 1287, pp.20-28.

Bartin, B., Ozbay, K., and Iyigun, C. (2006), A Clustered Based Methodology for Determining the Optimal Roadway Configuration of Detectors for Travel Time Estimation, Submitted to the 86th Transportation Research Board Annual Meeting.

Bertini, R.L. and Cassidy, M.J. (2002), Some Observed Queue Discharge Features at a Freeway Bottleneck Downstream of a Merge. Transportation Research Part A, 36, pp. 683-697.

Bianco, L., Confessore, G., and Reverberi, P. (2001), A network based model for traffic sensor location with implications on O/D matrix estimates, Transportation Science, Vol. 35, No. 1, pp. 50-60.

Cassidy, M.J. and Bertini, R.L. (1999) Some Traffic Features at Freeway Bottlenecks, Transportation Research Part B, 33, pp. 25-42.

Chen, C., Skabardonis, A., and Varaiya, P. (2003), Systematic Identification of Freeway Bottlenecks, 83rd Transportation Research Board.

Daganzo, C.F. (1997), Fundamentals of Transportation and Traffic Operations, Elsevier Science Inc, New York, pp. 133-135, 259.

Daganzo, C.F. (1994), The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory, Transportation Research Part B, Volume 28B, No. 4, pp. 269-287.

Daganzo, C.F. (1995), The Cell Transmission Model, Part II: Network Traffic, Transportation Research Part B, Volume 29B, No. 2, pp. 79-93.

Fei, X., Mahmassani, H.S., and Eisenman, S.M. (2007), Sensor Coverage and Location for Real-time Traffic Prediction in Large-Scale Networks, Transportation Research Record, Vol. 2039, pp. 1-15.

Fei, X. and Mahmassani, H.S. (2007), A Two-Stage Stochastic Model for the Sensor Location Problem in a Large-Scale Network, Conference Paper, 87th Annual Transportation Research Board Meeting.

Fujito, I., Margiotta, R., Huang, W., and Perez, W.A. (2005), The Effect of Sensor Spacing on Performance Measures.

Gendreau, M., Laporte, G., and Parent, I. (2000), Heuristics for the location of inspection stations on a network, Naval Research Logistics, Vol. 47, Issue 4, pp. 287-303.

Goldberg, D. (1989) Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, Inc., Reading, MA.

Hall, F.L. and Agyemang-Duah, K. (1991), Freeway Capacity Drop and the Definition of Capacity. Transportation Research Record No. 1320, pp. 91-98.

Hall, F.L. and Hall, L.M. (1991), Capacity and Speed-Flow Analysis of the Queen Elizabeth Way in Ontario, Transportation Research Record No. 1287, pp. 108-118.

Holland, J.H. (1975), Adaptation in Natural and Artificial Systems, University of Michigan Press (reprinted in 1992 by MIT Press, Cambridge, MA).

Koshi, M., Kuwahara, M., and Akahane, H. (1992), Capacity of Sags and Tunnels on Japanese Motorways, ITE Journal, May 1992, pp. 17-23.

Kwon, J., Petty, K., and Varaiya, P. (2006), Probe Vehicle Runs or Loop Detectors? Effect of Detector Spacing and Sample Size on the Accuracy of Freeway Congestion Monitoring, Submitted for Presentation and Publication at the Transportation Research Board – 86th Annual Meeting.

Newman, L. (1961), Study of Traffic Capacity and Delay at the Merge of the North Sacramento and Elvas Freeways, Report, California Division of Highways, USA.

Ozbay, K., Bartin, B., and Chien, S. (2004), South Jersey Real-Time Motorist Information Systems: Technology and Practice. Transportation Research Record, 1886, pp. 68-75.

Persaud, B.N. (1986), Study of a Freeway Bottleneck to Explore Some Unresolved Traffic Flow Issues, PhD Dissertation, University of Toronto, Canada.

Persaud, B.N. and Hurdle, V.F. (1991), Freeway Capacity: Definition and Measurement Issues, Proceeds of the International Symposium of Highway Capacity, A.A. Balkema Press, Germany, pp. 289-307.

Persaud, B.N., Yagar, S., and Brownlee R. (1998), Exploration of the Breakdown Phenomenon in Freeway Traffic, Transportation Research Record 1634, pp. 64-69.

Persaud, B., Yagar, S., Tsui, D., and Look H. (2001), Breakdown-related capacity for freeway with ramp metering, Transportation Research Record No. 1748, pp. 110-115.

Sherali, H.D., Desai, J., Rakha, H., and El-Shawarby, I. (2006), A Discrete Optimization Approach for Locating Automatic Vehicle Identification Readers for the Provision of Roadway Travel Times, Transportation Research Part B, Volume 40, Issue 10, pp. 857-871.

Yang, H., Yang, C., and Gan, L. (2006), Models and algorithms for the screen line-based traffic-counting location problems, Computers and Operations Research, Vol. 33, Issue 3, pp. 836-858.

Yang, H. and Zhou, J. (1998), Optimal traffic counting locations for origin-destination matrix estimation, Transportation Research Part B: Methodological, Vol. 32, Issue 2, pp. 109-126.

Zhang, L. and Levinson, D. (2004), Some Properties of Flows at Freeway Bottlenecks, Journal of the Transportation Research Board No. 1883, pp. 122-131.

Zhang, X. and Yang, H. (2004), The optimal cordon-based network congestion pricing problem, Transportation Research Part B, Vol. 38, Issue 6, pp. 517-537.

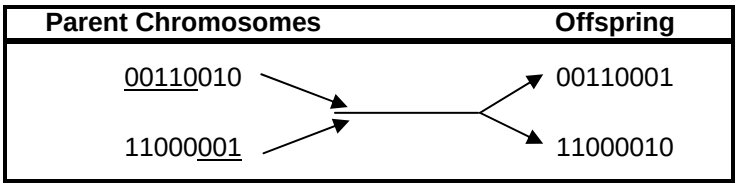


Table 1: Generation of offspring based on parental chromosomes

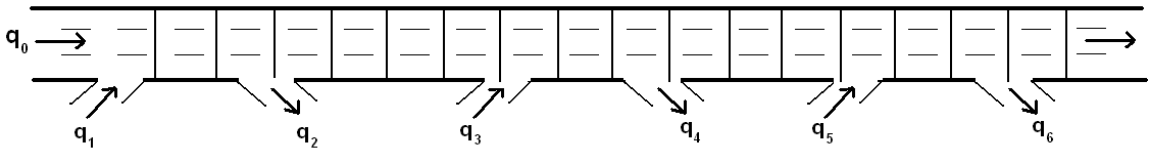


Figure 1: A freeway sketch, divided into cells

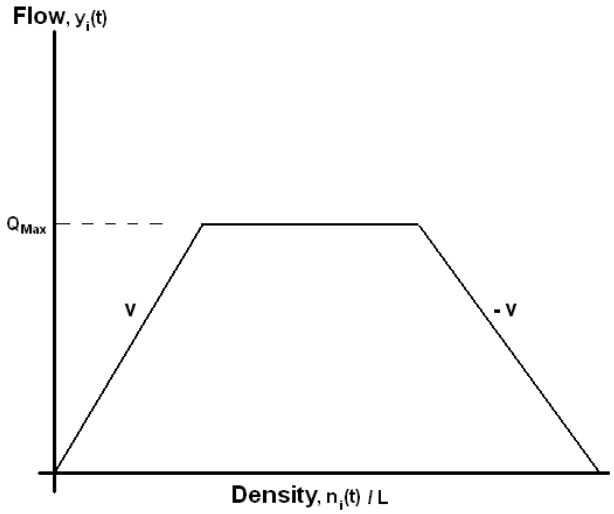


Figure 2: Trapezoidal flow-density diagram.

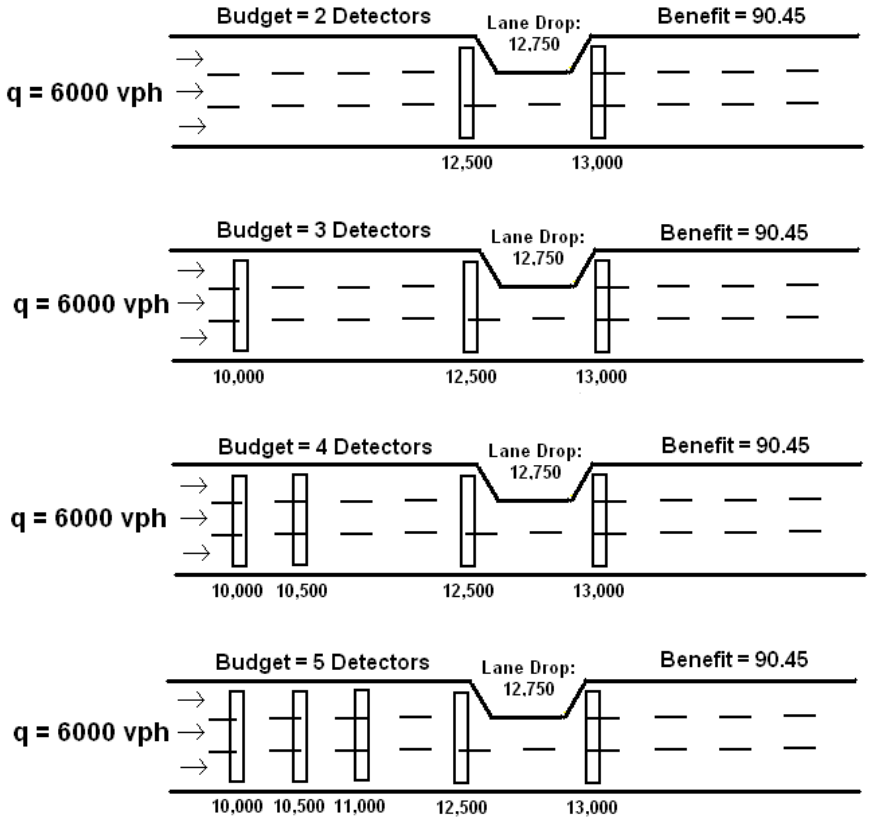


Figure 3: Placement of a designated number of loop detectors for a bottleneck. Traffic at this site moves from left to right. Stationing for each detector and the lane drop are in feet. Benefit is referred to the total benefit as claimed by the SP program for this configuration.

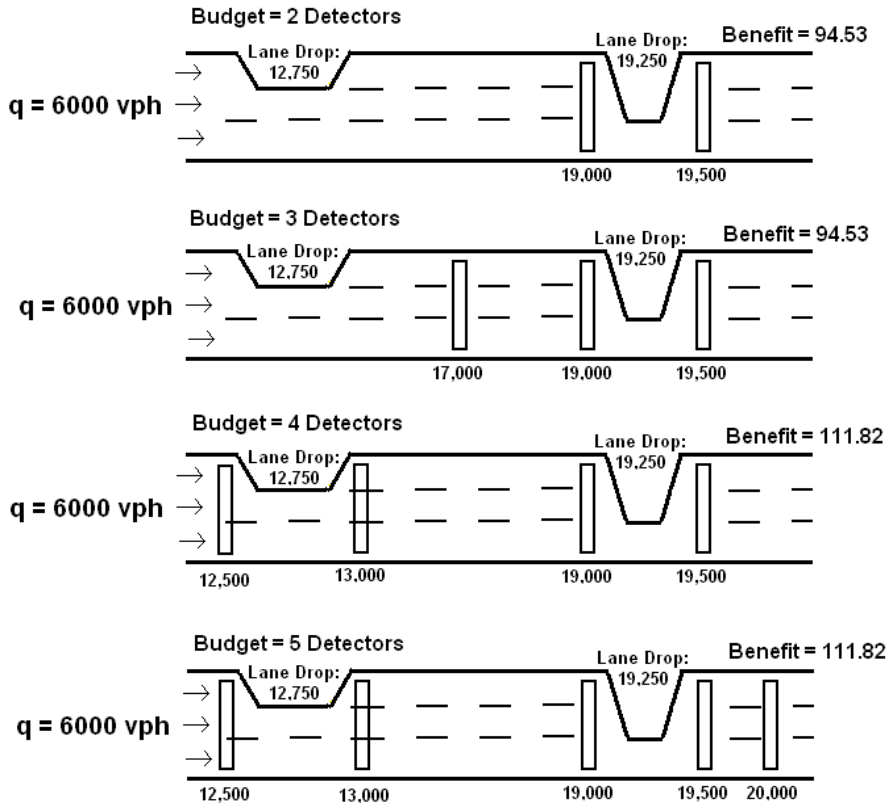


Figure 4: Placement of a designated number of loop detectors for a freeway with two bottlenecks.

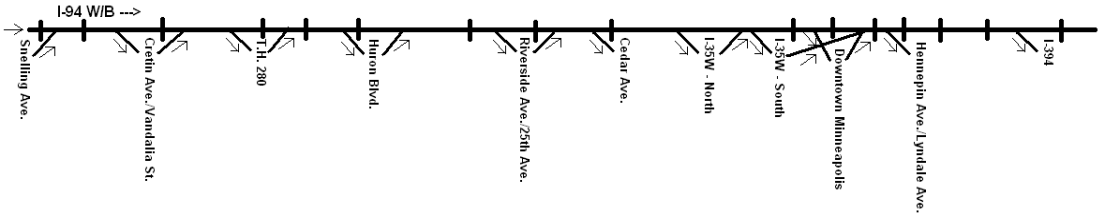


Figure 5: Existing loop detector configurations on the I-94 study site, not-to-scale. I-94 runs from Snelling Avenue in the east to I-394 in the west, totaling 7.2 miles. This study site passes through several busy areas, including the University of Minnesota, the I-35W interchange, and downtown Minneapolis. Dark vertical lines across I-94 represent loop detector placement where they are in the real world. 16 detectors are present.

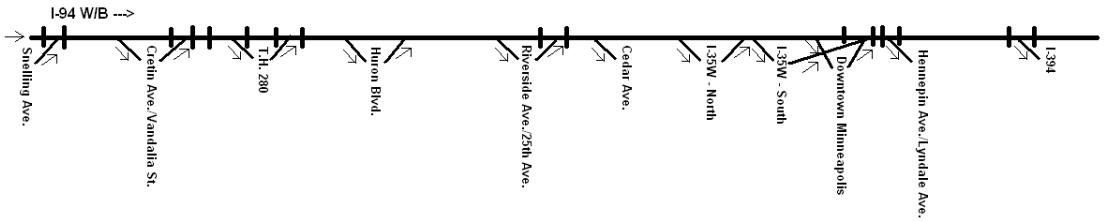


Figure 6: New loop detector configurations on the I-94 study site, not-to-scale. 16 loop detectors have been placed according to the SP program. Dark vertical lines represent these detectors.

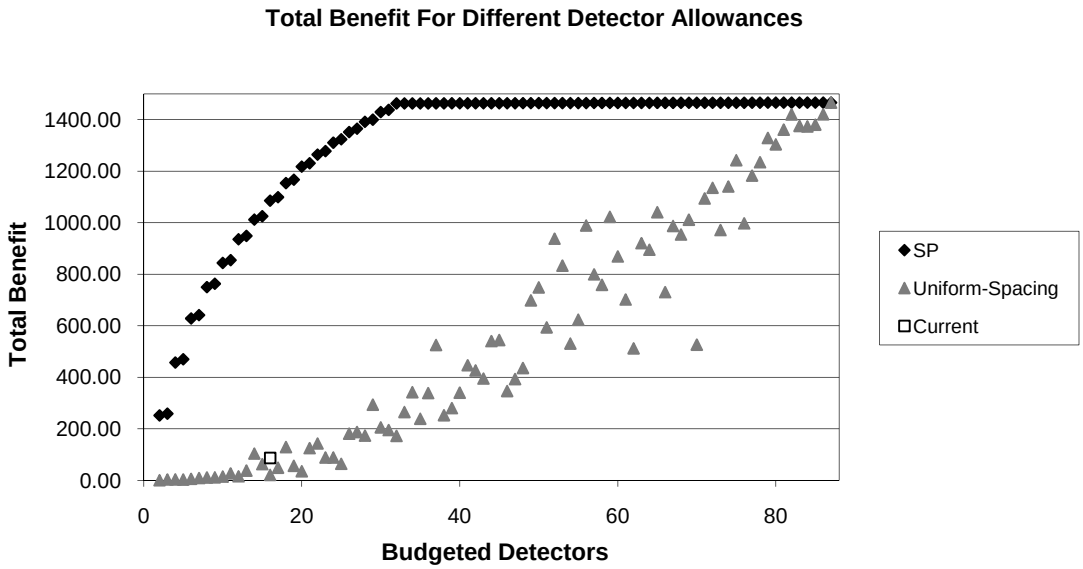


Figure 7: Total benefit for a varied number of allowable loop detectors.

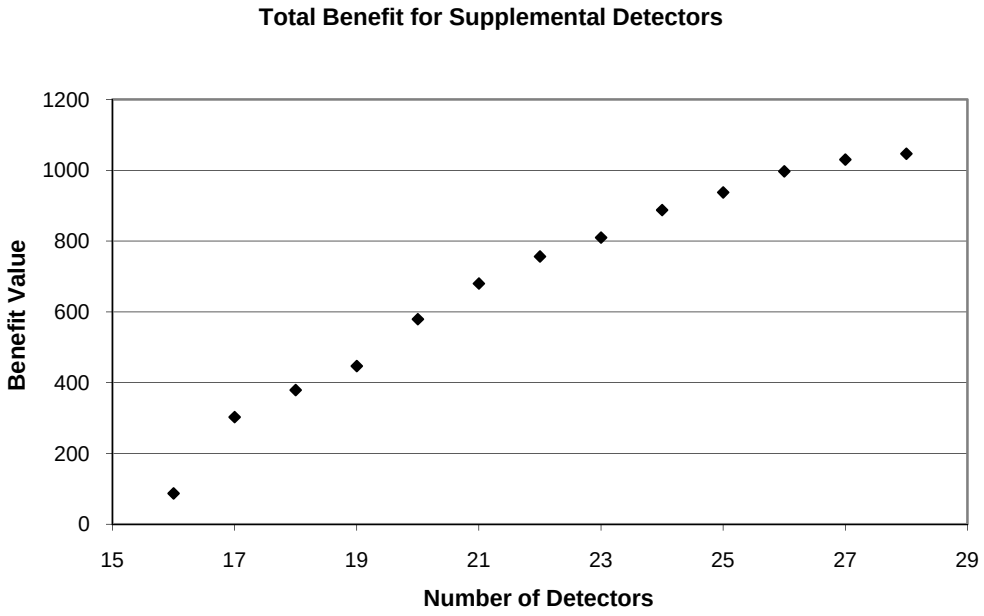


Figure 8: Total benefit when additional detectors are added, based on the SP program's recommendations, to an existing 16-detector corridor

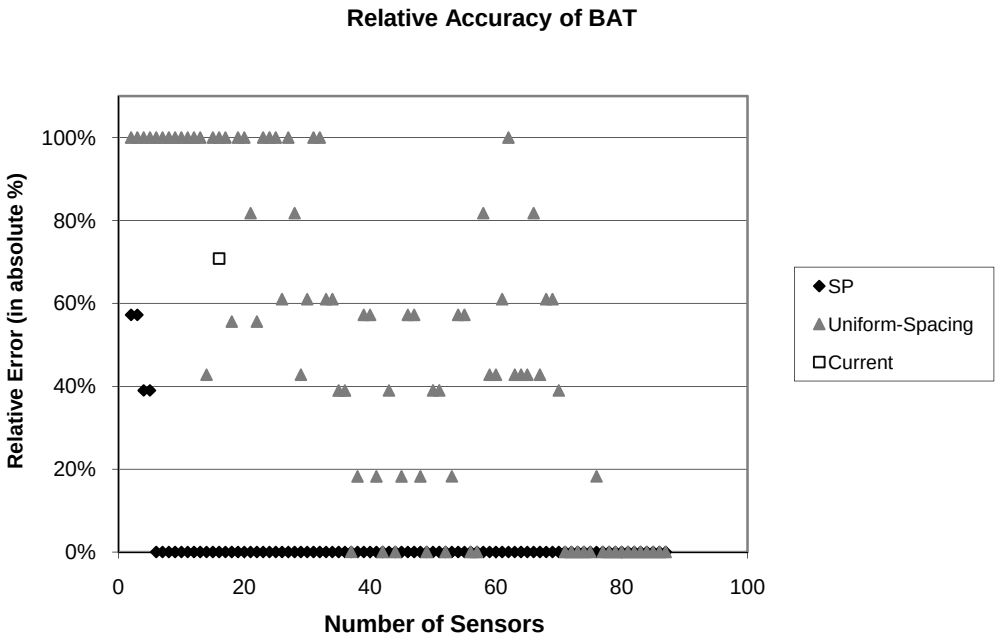


Figure 9: Relative error (in absolute percentages) between configurations for different sensor budgets when compared with ground-truth BAT

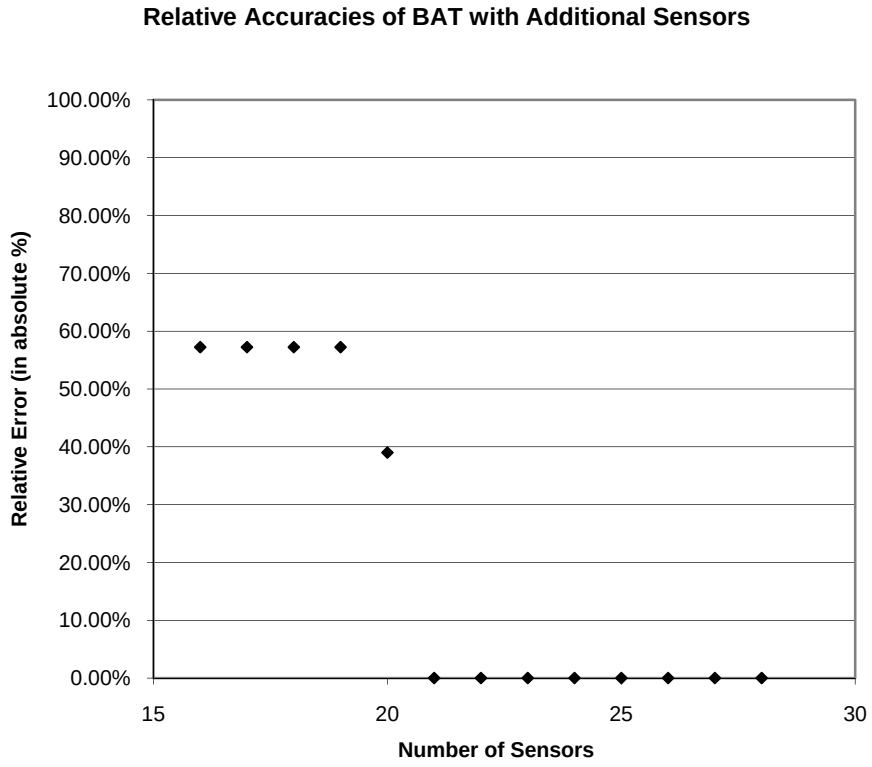


Figure 10: Relative error (in absolute percentages) for different sensor additions on an existing 16-detector corridor when compared with ground-truth BAT

SECTION 5 – ADVANCED STRATEGIES

X. BAN, L. CHU, R. HERRING, JD MARGULICI. “OPTIMAL SENSOR PLACEMENT CONSIDERING BOTH TRAFFIC CONTROL AND TRAVELER INFORMATION”. SUBMITTED TO 88TH ANNUAL TRANSPORTATION RESEARCH BOARD MEETING, 2008.

H. LIU, A. DANCZYK. “AN INTEGER LINEAR PROGRAM FOR OPTIMIZING SENSOR LOCATIONS ALONG CORRIDORS”. SUBMITTED TO TRANSPORTATION RESEARCH, PART B, 2008.

O.P. TOSSAVAINEN, R. HERRING, A. BAYEN. “FLOW MODEL BASED DENSITY ESTIMATION METHOD FOR HIGHWAYS USING OPTIMAL SENSOR CONFIGURATIONS”. CALIFORNIA CENTER FOR INNOVATIVE TRANSPORTATION, 2008.

Optimal Sensor Placement for Both Traffic Control and Traveler Information Applications

Xuegang (Jeff) Ban*
Department of Civil and Environmental Engineering
Rensselaer Polytechnic Institute (RPI)
110 Eighth Street, room JEC 4034
Troy, NY 12180-3590
Phone: (518) 276-8043 Fax: (518) 276-4833
Email: banx@rpi.edu

Lianyu Chu
CCIT, Institute of Transportation Studies (ITS)
University of California Berkeley
Tel: (949) 824-1876 Fax: (949) 824-8385
E-mail: lchu@berkeley.edu

Ryan Herring
Department of Industrial Engineering and Operations Research
University of California, Berkeley
Phone: 510-642-5667 Fax: 510-642-0910
Email: ryanherring@berkeley.edu

JD Margulici
CCIT, Institute of Transportation Studies (ITS)
University of California, Berkeley
2105 Bancroft Way, Suite 300
Berkeley, CA 94720-383
Phone: 510-642-5929 Fax: 510-642-0910
Email: jd@calccit.org

For Presentation and Publication
88th Annual Meeting of Transportation Research Board
August 01, 2008

# of Words: 5,846
+ (1 table and 7 figures) = 2,000
TOTAL: 7,846

* *Corresponding Author*

Abstract

Traffic sensors have been deployed for decades to freeways to meet the requirements of various traffic/transportation applications, most noticeably traffic control and traveler information applications. A unique feature of traffic sensor deployment is that it is a continuous and evolving process. That is, with new applications emerge, additional sensors are usually required to be deployed to meet new requirements. This process is also sequential in nature as the new deployment has to consider existing sensors. In this article, we propose a modeling framework to capture this sequential decision-making process for traffic sensor deployment. The framework is based on the Dynamic Programming (DP) model the authors recently developed for determining optimal sensor deployment for freeway travel time estimation. We illustrate the framework using two applications: ramp metering control and travel time estimation. It is found that the proposed scheme can appropriately capture the decision-making process of traffic sensor deployment, and can generate optimal sensor placement at any stage by considering sensors that have already been deployed in the field. The model is tested using GPS-enabled cell phone data on a real-world freeway route in the San Francisco Bay Area.

1 Introduction

Intelligent Transportation Systems (ITS) applications rely on various types of data (such as traffic flow, speed, or occupancy), which are usually collected through traffic sensors. For example, freeway travel time estimation often requires speeds measured at certain locations. Traditionally, many traffic sensors were deployed in a case by case base without a systematic study on where and how many sensors to deploy to fulfill the needs of the applications [1]. Since traffic sensors are limited (usually expensive) resources, determination of best deployment strategies can help maximize the value of this resource with minimum possible cost.

Recently, optimal sensor placement for providing traveler information especially travel times has received much attention. Most of the studies in this line focused on empirical investigations, i.e. by varying location and/or spacing of existing sensors, to study how sensor spacing impacts the travel time estimation quality [3, 4, 5, 6, 7, 8, 9]. Although empirical studies can provide some insights regarding information quality and the number of sensors, they could not provide reasoning why sensors should be deployed at certain locations. More rigorous modeling on optimal sensor placement can hopefully solve this issue, but is currently sparse in the literature. Sherali et al. [10] propose a mixed-integer optimization model to determine optimal placement of vehicle identification readers for travel time estimation, although the model can only be solved approximately. Bartin et al. [11] show that the optimal sensor placement for travel time estimation can be determined by minimizing the weighted summation of speed variations of all roadway segments, each of which is associated with a sensor. A nearest neighbor (NN) algorithm was further developed

[1] One exception is the optimal sensor location problem for origin-destination matrix estimation, which has been widely studied in the literature. See for example [1, 2] and references therein

in [11]. However, the NN algorithm is not guaranteed to provide a globally optimal solution in polynomial time.

A dynamic programming (DP) model and a shortest-path based solution algorithm are proposed in [12] to solve optimal sensor placement for freeway travel time estimation. The model is based on the observation that under certain conditions sensor deployment can be conducted in a staged process, in which the decision on one stage only depends on the starting state of that stage and not on any previous stages. The DP model requires availability of vehicle trajectories, which are not widely available in current practice. Therefore it is essentially an analysis framework. However, with the advent of GPS-enabled smart-phone-based traffic monitoring, the DP model can be used with probe vehicle data, which we will illustrate by using the Mobile Century data set, presented in the last part of this article. It is further shown that the DP model can solve large scale sensor deployment problems for travel time estimation to optimality in polynomial time.

Most previous studies on optimal sensor placement focuses on single applications only (e.g. travel time estimation). [2] In reality, sensors are seldom used for single purposes. Ideally one should consider all possible applications simultaneously and generate "optimal" sensor deployment to meet requirements of all these applications. However, this is highly impractical because 1) new applications always emerge and we cannot completely predict what will happen even for the near future, and 2) many sensors have already been deployed in the field for various applications, which we have to consider when deploying additional sensors for new applications. As a result, sensor deployment in reality works in a sequential manner with sensors deployed at different stages for different applications. One example of this is that freeway loop detectors were originally deployed for traffic control purposes, mainly ramp metering control. However, as the need to generate and disseminate traveler information emerged, they are now used (and may be supplemented by new sensors as well) to produce freeway travel time estimates.

In this article, we aim to model this sequential decision process. In particular, how can we make informed (or optimal) decisions on sensor deployment for certain application given some sensors have already been deployed? We illustrate the ideas using specific traffic control and traveler information applications. For this purpose, the answers to the following questions are crucial:

(1) Suppose we have a freeway route with a number of existing sensors. Are the existing sensors sufficient for traffic control purposes? If not, how to optimally supplement existing sensors to achieve the desired control goal?

(2) If the answers to (1) are affirmative, are the sensors sufficient for providing traveler information such as estimating travel times? If not, how to optimally supplement them to achieve desired quality of traveler information?

Notice that we group the questions as first for traffic control and then for traveler information applications. This is because we believe that these two types of applications have different priorities. At least from traffic management point of view, effective and efficient traffic management and control is the first priority. This is due to the following two reasons. First, historically traffic control and management is the focus of most traffic management agencies (like DOTs). This is true even when traveler information is becoming more crucial nowadays. Second and more importantly, a well managed and operated transportation system is more predictable and is thus the basis of effective traveler information. It is hard to imagine that traveler information will have significant value on a poorly managed transportation system. Therefore, we need to solve sensor placement for traffic control applications first. This is actually what is happening now: there

[2] One exception is Eisenman et al. [13] who provide an information learning based conceptual framework of sensor placement for various applications. But the framework is too general to be practically implementable.

are already some sensors in place (most likely for some traffic control purposes) and we need to optimally enhance these sensors for new applications (e.g. traveler information applications).

In this article, we try to answer the above two questions in a sequential yet coherent manner. In particular, we focus on two applications: freeway ramp metering control and providing freeway travel times. The key is the modeling ability to optimally add additional sensors to the field if needed to meet the requirements of new applications. In this article, we show that the DP model developed in [12] can be extended to determine optimal sensor placement for other applications such as occupancy estimation. Furthermore, it is able to consider existing sensors when determining the optimal locations of new sensors. In Section 2, the DP model and solution algorithm is briefly described, together with the method of how to optimally incorporate existing sensors. The ramp metering control application is discussed in Section 3. In this article, we focus on SWARM [14] which requires freeway mainline occupancy as the major input. SWARM also requires sensors at fixed locations (usually upstream) for each on-ramp, and we will then show how to determine the optimal locations of additional sensors to have appropriate estimation of freeway occupancy. Travel time application is presented in Section 4. Section 5 focuses on numerical examples based on a segment of I-880 in the San Francisco Bay Area. The data are obtained from a field experiment which deployed for 10 hours 100 cars equipped with GPS-enabled cellular phones to collect traffic data along the segment of freeway. We conclude our work in Section 6.

2 Dynamic Programming Model for Optimal Sensor Placement

2.1 A DP Model

Denote the study freeway segment as route r with length L. Assume the duration of the study period is T. WSe discretize time into *intervals* and space into *sections*. Each interval has fixed duration Δt such as 30 seconds; each section has also fixed length Δx such as 50 or 100 feet. We assume $L = N\Delta x$ and $T = H\Delta t$, i.e. we have in total N sections and H time intervals. Given this setting, suppose we are interested in some generic traffic measurement which could be for example speed, occupancy, etc. In particular, we denote the *ground-truth* measurement at time t ($1 \le t \le H$) and section n ($1 \le n \le N$) as $u(n,t)$. Assume we are to deploy K sensors to route r and in general $K \ll N$. Further denote the *estimated* measurement at time t and section n is $\bar{u}(n,t)$, which is generated by the sensors. Here we assume that sensors should be deployed to minimize the deviation between the ground-truth and estimated measurements. A commonly used metric is the mean square error (MSE), defined as follows:

$$E = \frac{\sum_{n=1}^{N}\sum_{t=1}^{H}(u(n,t) - \bar{u}(n,t))^2}{NH}. \tag{1}$$

In this article, E is the objective function that will be used to determine the optimal sensor placement. Now the question is how the estimated measurements are obtained given sensor placement. We adopt a simple yet practical scheme: each sensor is associated with a spatial *influence area*, called a link. Each link contains one or multiple sections and the link boundaries are the section boundaries. We assume the estimated measurement of every section within a link is identical to the measurement collected by the sensor associated with the link. This is illustrated in Figure 1, which depicts a route with 7 sections numbered 1 - 7. The three sensors are denoted as I, II, III. Sensor I is associated with a link that contains sections 1-3, the link of sensor II contains section 4, and sections 5-7 are for the link associated with sensor III. The ground-truth

and estimated measurements for a given time t are shown using two curves at the top of Figure 1. Notice that since sensor I is deployed at section 2, the sensor-generated measurement by I is $u(2,t)$ for any time interval t. In other words, we assume in this article that sensors are "perfect" in detecting the measurement. As a result, the estimated measurements for the first three sections are $\bar{u}(i,t) = u(2,t), \forall 1 \leq i \leq 3$. Similarly, we have $\bar{u}(4,t) = u(4,t)$ and $\bar{u}(i,t) = u(6,t), \forall 5 \leq i \leq 7$. That is, the estimated measurement at one section at time t is the ground-truth measurement at another section (maybe itself) at the same time. The resulting estimated measurements are represented using the step-wise curve since we assume the sensor measurement is uniform across its associated link. The optimal sensor placement should then minimize the deviation of the two curves over all time intervals.

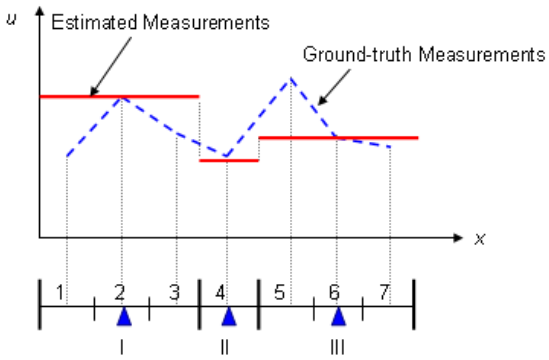


Figure 1: Calculation of Estimated Measurements

If we further assume that a sensor is always in the middle of its associated link, we can then convert the optimal sensor placement problem to a problem that aims to determine the optimal link starting and ending locations. Clearly the starting and ending locations of link k $(1 \leq k \leq K)$ coincide with sections, denoted as s_k and y_k respectively. The objective function E in (1) can then be rewritten as:

$$E = \frac{\sum_{k=1}^{K} \sum_{s_k \leq n \leq y_k} \sum_{t=1}^{H} (u(n,t) - \bar{u}(n,t))^2}{NH} = \sum_{k=1}^{K} E_k(s_k, y_k). \tag{2}$$

In the above equation, E_k is the MSE of measurements for link k which is a function of the starting and ending link locations s_k, y_k only. Specifically, E_k can be defined as:

$$E_k(s_k, y_k) = \frac{\sum_{s_k \leq n \leq y_k} \sum_{t=1}^{H} (u(n,t) - \bar{u}(n,t))^2}{NH}. \tag{3}$$

We assume a set of vehicle trajectories are available for the study route from which $u(n,t)$ can be estimated for any (n,t) pair. Under this assumption, $E_k(s_k, y_k)$ is computable for any given starting and ending location s_k, y_k. As a result, the optimal values of $s_k, y_k, 1 \leq k \leq K$ can be obtained via solving the following integer programming problem:

$$\min_{1\le s_k, y_k \le N, k=1,\ldots,K} \quad \sum_{k=1}^{K} E_k(s_k, y_k). \tag{4}$$

$$\text{s.t.} \qquad s_1 = 1 \tag{5}$$

$$y_K = N. \tag{6}$$

$$s_{k+1} = y_k + 1. \tag{7}$$

$$k \le s_k \le y_k \le N - K + k. \tag{8}$$

Here (5) - (8) are constraints for sensor placement. Equations (5) and (6) hold because the first link must start at section 1 and the last link (link K) must end at section N. Equation (7) holds since knowing the ending section of link k (y_k), the starting section of link $(k+1)$ must be the next section $(y_k + 1)$. This is called *state transfer*. The first inequality of (8) holds since there are $k-1$ links before link k, which contain at least $k-1$ sections. Similarly, the last inequality of (8) holds since there are $K-k$ links after link k, which contain at least $K-k$ sections.

Solving the above integer programming problem for large scale problems is not tractable. In this article, we convert the problem to a DP model by 1) dividing the problem into K stages (one sensor is deployed in each stage), 2) defining s_k as the state variable and y_k as the decision variable of stage k, and 3) assuming the cost at each stage is the link MSE (E_k as defined in (3)). Under this setting, it is clear to see that the decision at stage k (i.e. the value of y_k) is only determined by the state variable of the stage (i.e. s_k) since the objective to deploy sensor k at this stage is to minimize E_k which is a function of s_k and y_k only. This observation leads to a DP formulation of the problem. Further define $F_k(s_k)$ as the total cost from stage k (including stage k) to the last stage (i.e. stage K). Then a recursive formulation exists for $F_k(s_k)$ as follows:

$$F_1(s_1) = F_1(1) = \min_{1\le y_1 \le N-K+1} \left\{ E_1(1, y_1) + F_2(y_1 + 1) \right\}, \tag{9}$$

$$F_k(s_k) = \min_{s_k \le y_k \le N-K+k} \left\{ E_k(s_k, y_k) + F_{k+1}(y_k + 1) \right\}, 2 \le k \le K-1, \tag{10}$$

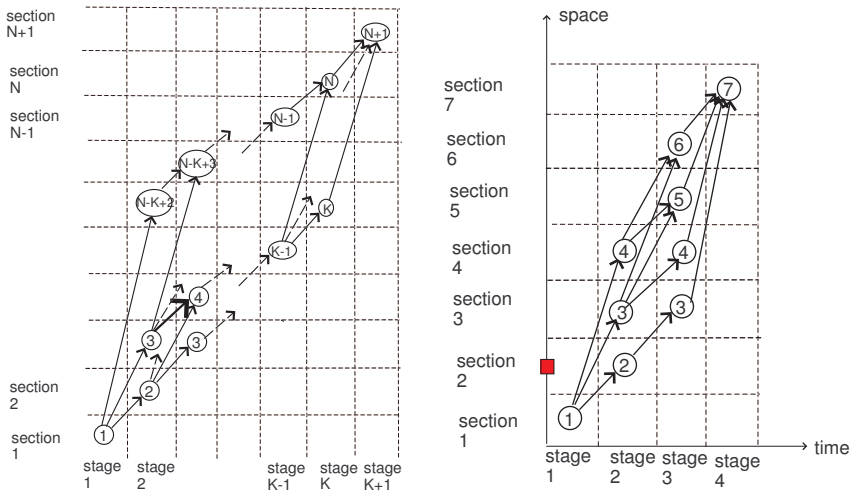
$$F_K(s_K) = E_k(s_K, N). \tag{11}$$

It can be shown that solving (4) - (8) is equivalent to solve these three recursive equations which satisfy the *optimality principle* of DP [12].

2.2 A Graph-Based Solution Algorithm

The DP model (9) - (11) can be represented as a graph shown in Figure 2(a). In the figure, stages are listed horizontally and sections are listed vertically. The state of a stage represents the starting section of the link associated with the stage. In this figure, all possible states of a stage are represented as *nodes*. In other words, a node represents a section of the roadway, and the node number is the section number. For example, the node at stage 2 and Section 2 represents that the starting location of link 2 could be section 2. As mentioned before (especially equations (5) - (8)), there is only one state in stage 1 ($s_1 = 1$) and $(N - K + 1)$ states (from k to $N - K + k$) for stage $k = 2, \ldots, K$. We further create a fake stage as stage $K + 1$ that has only one fake state $N + 1$.

A connection, denoted as an *arc*, may be created from a node in stage k to another node in the immediate next stage $k + 1$ if the latter node has a higher node number. Each arc actually represents a possible roadway link by defining the link's starting and ending sections. That is, an



(a) Graph Representation of DP Model. (b) Consideration of Existing Sensors

Figure 2: The DP Model

arc from node s_k in stage k to node s_{k+1} in stage $k+1$ represents one possible configuration for link k: it starts at section s_k and ends at section $s_{k+1}-1$ because the next link starts at s_{k+1}. Therefore, we must have $s_{k+1} > s_k$ in order to construct the arc. For example, the arc from node 2 in stage 2 to node 4 in stage 3 (marked in bold line) in Figure 2(a) means that one possible configuration for link 2: it starts at node 2 and ends at node 3 (both are inclusive). Therefore, there should be no arc from node 4 in stage 2 to node 4 or lower in stage 3. Furthermore, there are no arcs between any two stages that are not adjacent to each other. We also associate a cost with each arc in Figure 2(a). For the arc from node s_k in stage k to node s_{k+1} in stage $k+1$, the arc cost is $E_k(s_k, s_{k+1}-1)$ as computed in (3). In other words, the cost of an arc is the MSE of travel time estimation for its corresponding roadway link.

It is easy to check that the graph constructed in the above manner enumerates all possible states in each stage (1 to K) and all possible configurations (i.e., the starting and ending locations) of each link. It also incorporates all the constraints of the model shown in equations (5) - (8). More importantly, each path from node 1 in stage 1 to node $N+1$ in stage $K+1$ contains exactly K arcs, each of which represents a possible configuration of a particular roadway link (i.e. its starting and ending sections). In other words, each path represents a potential sensor deployment scenario. Therefore the optimal sensor locations can be achieved by finding the minimum-cost path from node 1 in stage 1 to node $N+1$ in stage $K+1$. Since all arc costs are positive, the DP model can be solved by a shortest-path search algorithm. Furthermore, the time complexity of solving the DP is $O(KN^2)$ if $N \gg K \gg 2$, which is polynomial.

2.3 Consideration of Existing Sensors

One feature of the DP model and its graph-based solution approach is that existing sensors can be easily considered. In this case, we make a simple adjustment to the dynamic programming graph representation of the solution space. First, we match all existing sensors to the appropriate section they reside in. Then, every possible link (represented as an arc in the graph) that covers a section with an existing sensor in it but does not have the existing sensor at the center of the link

is removed from consideration as a possible choice in the final solution. The reason for this is that we assume a sensor must be in the middle of its associated link.

As an example, Figure 2(b) shows a highway segment that is broken down into 6 sections. Suppose that we already have a sensor in section 2. If this is the case, then we cannot consider links that cover section 2 but do not have section 2 as the middle of the link. This means that a link covering sections 1 through 4 would not be permissible in the solution (because that would imply a sensor in 3 and not on section 2 based on the fact that a sensor must be in the middle of its associated link). This also implies that the arc from node 1 in stage 1 to node 5 in stage 2 in the DP graph should be eliminated. On the other hand, a link covering sections 1 through 3 would be permissible, implying that the arc from node 1 in stage 1 to node 4 in stage 2 should be included. This is also the case for the arc from node 1 in stage 1 to node 2 in stage 2, and the arc from node 1 in stage 1 to node 3 in stage 2. Furthermore, the arcs from node 2 in stage 2 to nodes 4, 5, and 6 in stage 3 should all be eliminated. The graph in Figure 2(b) shows the adjusted DP graph after removing all impermissible links.

As a result, to account for existing sensors, one can use a simple linear search on all of the links to identify which ones to remove, and then uses the shortest path algorithm described in Section 2.2 to compute the final solution on the adjusted graph. Therefore, the complexity of the algorithm remains the same as the original DP algorithm, i.e. $O(KN^2)$ for $N \gg K \gg 2$. The DP model and graph-based solution technique presented in this section is general and may be applied to freeway speeds, occupancies, and travel times. The only difference for different applications is how the objective function is defined, especially the link MSE E_k as defined in (3). In the next two sections, we discuss how the model can be used for ramp metering and freeway travel time applications.

3 Optimal Sensor Placement for Ramp Metering

3.1 Ramp Metering Background and History

Ramp metering has been recognized as an effective freeway management strategy to avoid or ameliorate freeway traffic congestion by limiting access to the freeway. The benefits of ramp metering are:

(1) Restrict vehicles entering freeway by temporarily storing them on the ramps in order to ensure that mainline freeway is operated within capacity and thus prevent congestion.

(2) Break up platoons of vehicles entering freeways in order for vehicles from onramps to merge to the freeway mainline more easily and thus enhance safety.

(3) Divert vehicles that cannot afford waiting on the onramps to other routes and thus reduce demand to the freeway.

In practice, methods of metering operation can be divided into two primary categories: fixed-time (or pre-timed) control and adaptive (or traffic responsive) control. Pre-timed control utilizes Time-of-Day metering rates that are pre-determined to best manage “expected” conditions based on an analysis of historical data. Adaptive control dynamically modifies metering rates based on real-time traffic data, thus, conceptually allowing for better responses to variations in traffic conditions.

The adaptive or traffic responsive ramp metering control can be further classified as local traffic responsive control and coordinated traffic responsive control. Local traffic responsive control determines metering rates based on current prevailing mainline traffic conditions in the vicinity of the ramp. Examples are demand-capacity control, occupancy control, and feedback control [15]. Coordinated traffic responsive control determines metering rates based on the prevailing traffic conditions of an extended section of roadway. Notable instances include ZONE in Minnesota, BOTTLENECK in Washington, and SWARM in California and Oregon [16, 17].

3.2 Sensor and Data Requirement for Ramp Metering

Local traffic-responsive ramp metering control needs to obtain traffic condition data from sensors on the freeway mainline. Typically, these sensors are required to be placed upstream of the ramp. This requirement applies to many ramp metering systems deployed in the real world, such as the three metering systems in California: Semi-Actuated Traffic Management System (SATMS), San Diego Ramp Metering System (SDRMS), and Traffic Operations System (TOS). All three systems need mainline traffic flow and occupancy data to operate.

Coordinated traffic-responsive ramp metering control seeks to optimize a multiple-ramp section of a highway, often with the control of flow or occupancy through a bottleneck as the ultimate goal. Typically, sensors are needed to be placed to the mainline freeway evenly at a certain space and/or at bottleneck locations. For example, ZONE needs volume data; BOTTLENECK needs occupancy data; SWARM needs either volume and/or occupancy data.

By looking at the type of data requirement, it is found that occupancy data are widely used and required by most ramp metering algorithms. In this article, we focus on SWARM and investigate how sensors can be deployed to better facilitate this metering algorithm.

3.3 Optimal Sensor Placement for SWARM

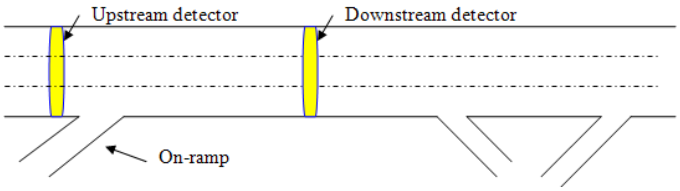
SWARM is a system-level ramp metering system. It is operated as a central ramp metering system at Traffic Management Center (TMC). SWARM has four algorithms, SWARM 1, SWARM 2a, SWARM 2b, and SWARM 2c. A meter operated under SWARM can be set up to use either of them or the combination of them. SWARM needs a mainline sensors to be placed upstream of each onramp. In order for SWARM 2b and 2c to work appropriately, as shown in Figure 3(a), SWARM requires a sensor located upstream of the ramp and another located downstream of the ramp. The mainline sensor upstream of the next onramp can be used as the "downstream" sensor, as shown in Figure 3(b). However, this relaxation will not work well if thee is a bottleneck, caused by either lane drop or strong weaving or merging, between the two mainline upstream sensor. As a result, we assume that the upstream sensors of all onramps as given (since they have to be deployed as a requirement by the algorithm); the downstream sensors however need to be "optimally" determined in terms of both numbers and actual locations to better estimate the mainline occupancy. As discussed in Section 2, the objective is to minimize the deviation of ground-truth and estimated mainline occupancy, as defined as follows:

$$E_c = \sum_{k=1}^{K} E_k^c(s_k, y_k). \tag{12}$$

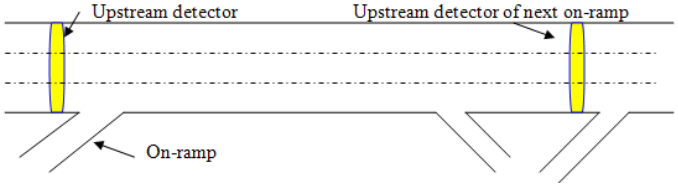
Here E_c denotes the objective function for ramp metering (i.e. for occupancy) and E_k^c is the MSE of occupancy for link k which can be defined as:

$$E_k^c(s_k, y_k) = \frac{\sum_{s_k \le n \le y_k} \sum_{t=1}^{H} (o(n,t) - \bar{o}(n,t))^2}{NH}. \quad (13)$$

We use $o(n,t)$ and $\bar{o}(n,t)$ to denote respectively the ground truth and estimated occupancies at section n and time t.



(a) Normal Sensor requirement for SWARM 2c



(b) Relaxed Sensor requirement for SWARM 2c

Figure 3: Sensor Placement Requirement for SWARM Ramp Metering Algorithm

4 Optimal Sensor Placement for Freeway Travel Time Estimation

Detailed discussions on how to apply the DP model to determine optimal sensor deployment for freeway travel time estimation are provided in [12]. In this section, we briefly discuss how the objective function is defined. For this purpose, we first denote respectively $\hat{\tau}_k^m$ and τ_k^m the estimated and actual travel times of the m-th vehicle ($1 \le m \le M$) traveling link k ($1 \le k \le K$), and M is the total number of vehicles. The travel time estimation error for the m-th vehicle on link k, denoted as e_k^m, can be expressed as:

$$e_k^m = \hat{\tau}_k^m - \tau_k^m. \quad (14)$$

We then use the same objective function as that in [11, 12], which is defined as follows:

$$E_t = \frac{\sum_{m=1}^{M} \sum_{k=1}^{K} (e_k^m)^2}{M} = \sum_{k=1}^{K} E_k^t. \quad (15)$$

Here E_t represents the objective function for travel time estimation. E_k^t is the Mean Square Error (MSE) of the travel time estimation for all M vehicles for link k, defined as:

$$E_k^t = \frac{\sum_{m=1}^{M} (e_k^m)^2}{M}. \quad (16)$$

The objective defined in (15) focuses on estimation errors of all individual links, instead of only on the entire route. The reason for this is that we want to generate sensor locations that can provide "good" estimates for all link travel times, not only in terms of the entire route. If attention is only put on the entire route, it is possible that the resulting sensor locations may underestimate travel times for certain links and overestimate for other links; but as a whole, they cancel out each other and provide good estimation. This type of sensor placement is not desirable. It is easy to see that the objective function we use here can effectively eliminate such sensor deployment strategies since they will lead to large objective values using equation (15). In [12], it is shown that this objective function definition is effective to generate sensor placement that is optimal to both the (entire) route r and its sub-routes.

5 Case Studies

We present a case study in this section to illustrate how the optimal sensor placement can be determined by considering ramp metering and travel time estimation in a sequential manner. The case study is based on Mobile Century data. Mobile Century is an experiment performed on February 8th, 2008, in which 165 drivers drove 100 vehicles on Interstate 880 (see Figure 4) for 10 hours in loops of length 5.5 to 10 miles [18, 19]. The experiment involved each vehicle carrying a Nokia N95 GPS-enabled smartphone, transmitting in real time loop detector-like data (called VTL data for Virtual Trip Line), which consists of speed readings at GPS-defined locations upon crossing of the locations. These VTLs represent "virtual" loop detectors, which are smartphone-based and may be used by phone manufacturers and access providers to monitor traffic in the near future. The experiment achieved a 2% to 5% penetration rate on the highway throughout the day, thus mimicking smartphone penetration in the driving population in about 18 months. In addition to this online transmitted data, each of the GPS logs collected by the phones at a 1/3 Hz rate was saved in the memory of the phone. While using trajectory data is not part of the Mobile Century technology development plan, this archival data collected from the experiment can be of great use for traffic modeling and analysis (as will be shown later in this section).

We select the shorter loop from CA-84 (postmile 20) to Tennyson St (post mile 25.5) as the study site in this article. The two circles in Figure 4 show the starting and ending locations of the route, which is about 5.5 miles. There are five major interchanges along this route at Decoto Rd, Alvarado Blvd, Alvarado Niles Rd, Wipple Rd, and Industrial Pkwy respectively. All the on-ramps to I-880 at these five locations are metered. The currently deployed ramp metering strategy is TOS, which is local responsive. TOS may be upgraded in the future to system-wide metering strategies. In this article, we select SWARM as one alternative for system-wide metering scheme. Since upstream sensors are required by SWARM (see Section 3.3), we consider the nearest upstream mainline sensor at each major interchange as existing sensors. Their locations are marked using solid lines in Figure 4 together with their exact location in postmile.

If we evaluate equation (2) for the five existing sensors for occupancy, the resulting objective value (we take the square root of (2) hereafter in this article so that the objective value has the same unit as the traffic measurement, i.e. occupancy or travel time) is 0.0264 (2.64%). This implies that the average deviation between the ground-truth occupancy (calculated using vehicle trajectories from Mobile Century) and the estimated occupancy (estimated using the scheme in Section 2.1) is 2.64%. For illustration purpose, suppose we require the deviation must be less than 0.0225 (2.25%). We will then need to add more sensors to this route. As shown in Section 2.3, the optimal placement of additional sensors can be determined using the DP model. The result shows that adding 4 additional sensors will service the purpose and the deviation is reduced to 0.0224 (2.24%). Figure 5 depicts (using the line marked as "Occ") how the occupancy objective

Figure 4: Study Site of the Case Study (Source: maps.google.com)

value changes as the number of sensors increases from 5 to 9. The reduction of the objective value is monotonic (meaning the estimation quality is improved with more sensors deployed) which is a desirable feature of the DP algorithm.

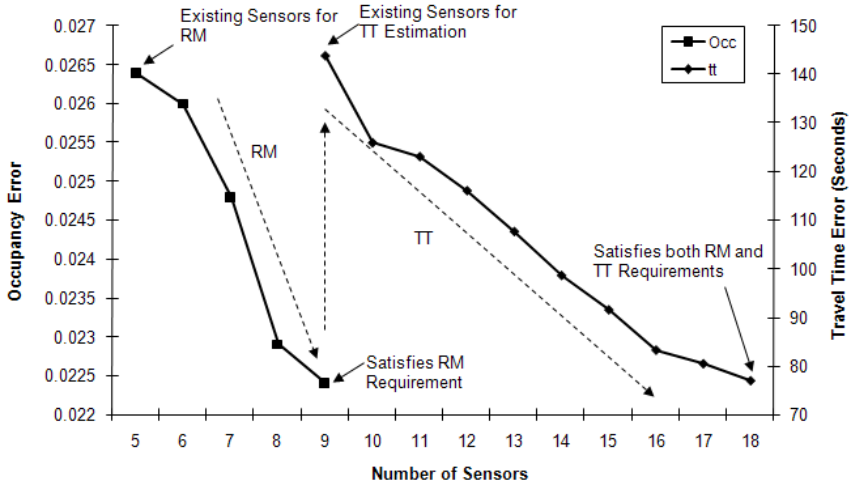


Figure 5: Change of Objective Values vs. Number of Sensors

After satisfying the ramp metering application requirement, we now focus on the travel time estimation application. The nine sensors produce an objective value of 143.8 seconds based on

equation (16). If we require the objective value for travel time must be less than 80 seconds (which represents about 13% error since the travel time of the entire route is about 10 minutes, i.e. 600 seconds). It turns out that 9 additional sensors are needed, resulting in 18 sensors in total for this segment. The line marked with "tt" in Figure 5 illustrates how the travel time objective value changes as the number of sensors increases from 9 (as a result of ramp metering requirement) to 18. Again, the decrease of the travel time estimation error is monotonic. We show the actual locations of the 18 sensors in Table 1. The second column of the table lists the exact postmiles of the 18 sensors, while columns 3-5 indicate using "√" whether a particular sensor is considered as existing or generated for ramp metering or for travel time estimation.

Sensor Index	PM	Existing Sensors	Ramp Metering	Travel Time
1	20.51	√		
2	21.28	√		
3	22.1		√	
4	22.98			√
5	23.33	√		
6	23.68			√
7	24.01	√		
8	24.12			√
9	24.24		√	
10	24.31			√
11	24.4			√
12	24.48	√		
13	24.57			√
14	24.71			√
15	24.8		√	
16	25.01			√
17	25.3		√	
18	25.45			√

Table 1: Optimal Sensor Locations for the Case Study

The dashed arrows in Figure 5 depict how the objective values for the two applications change as the number of sensors increases. Notice that we first focus on ramp metering application (as marked as "RM") which has higher priority. As we reach the stage when we have 9 sensors, the ramp metering requirement is met. This is the time when we switch our focus to travel time estimation. The vertical dashed arrow indicates this switch and the 9 sensors from the ramp metering application is the "existing" sensors for travel time estimation. After the switch, we concentrate on the travel time application as marked by "TT." Notice that finally we deploy 18 sensors, which satisfy the requirements for both ramp metering and travel time estimation applications. It is the authors' understanding that the above decision-making process is more closely related to what is happening in practice.

To show whether the generated sensor placement makes sense, we associate the sensor locations with the speed contour map of the route. Figure 6 depicts the speed contour map of the route with darker color representing more congested areas. It is clear that the major congestion area is roughly from PM 23.5 to 25.5. The triangles on the y-axis of the figure shows the sensor locations for 9 sensors. By comparing the locations of existing five sensors in Figure 4 (also shown in Table 1), we can see that only 1 additional sensor is deployed to the free flow area (at PM 22.1) and the other 3 sensors are all deployed to the congestion area (at PM 24.24, 24.8, and 25.3 respectively). This intuitively makes sense since traffic conditions at congestion areas are usually more complicated which need to be captured by additional sensors.

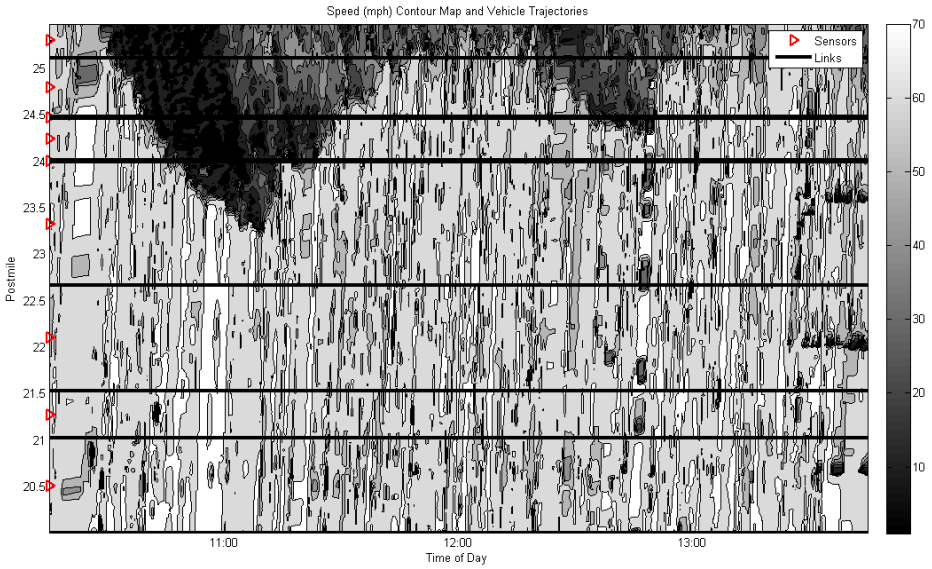


Figure 6: Optimal Sensor Placement vs. Speed Contour Map (9 Sensors)

We further depict in Figure 7 the generated sensor locations by the DP model as we go through the entire process, i.e. by increasing the number of sensors from 5 (the beginning state) to 18 (which satisfies both the ramp metering and travel time estimation). In the figure, asterisks represent the locations of sensors. We can see that as more sensors are deployed, most of them are deployed to the congestion area (PM 23.5 to 25.5); only 1 (up to 11 sensors in total) or 2 (12 sensors or above in total) additional sensors are deployed into the free flow area. More importantly, as additional sensors are added in, the locations of previously deployed sensors in congestion areas remain almost unchanged. This is illustrated using the solid thin lines in the figure (dashed lines indicate the locations of the five existing sensors), which show that locations of newly deployed sensors just "branch out" from existing sensors in congestion areas. This implies that the DP algorithm has the ability to capture the most significant congestion area and if more sensors are available, the second most significant congestion area will be captured and so on. The locations of sensors in free flow areas however may change since the speeds detected in free flow areas are not sensitive to the actual sensor locations. The above discussions illustrate the close correlation between the optimal sensor locations generated by the DP algorithm and the congestion areas of the network. They also show that the results from DP are stable and predictable, which is desirable in practice.

To illustrate that the generated optimal sensor placement is superior to that generated by purely engineering judgement, we compare the performance of the model-generated sensor placement with sensors that have already deployed in the field along the study route. As shown on PeMS (Performance Measurement Systems, [20]), there are 13 sensors deployed on I-880 NB for the study route (i.e. from PM 20 to 25.5). By evaluating the objective value of these 13 sensors using equation (15), we can obtain that the estimation error is about 159.50 seconds, roughly 26.6% (the route travel time is 10 minutes). Now we assume we have these 13 sensors but will place them at different locations based on the DP model. This results in an estimation error of 107.70 seconds, approximately 18.0%, which can be seen from Figure 5. In other words, the modeling framework proposed in this article could potentially improve the travel time estimation quality by 8.6% by having installed those sensors at more appropriate locations, and without any other extra cost. We notice that this comparison only considers ramp metering and travel time estimation. However, it

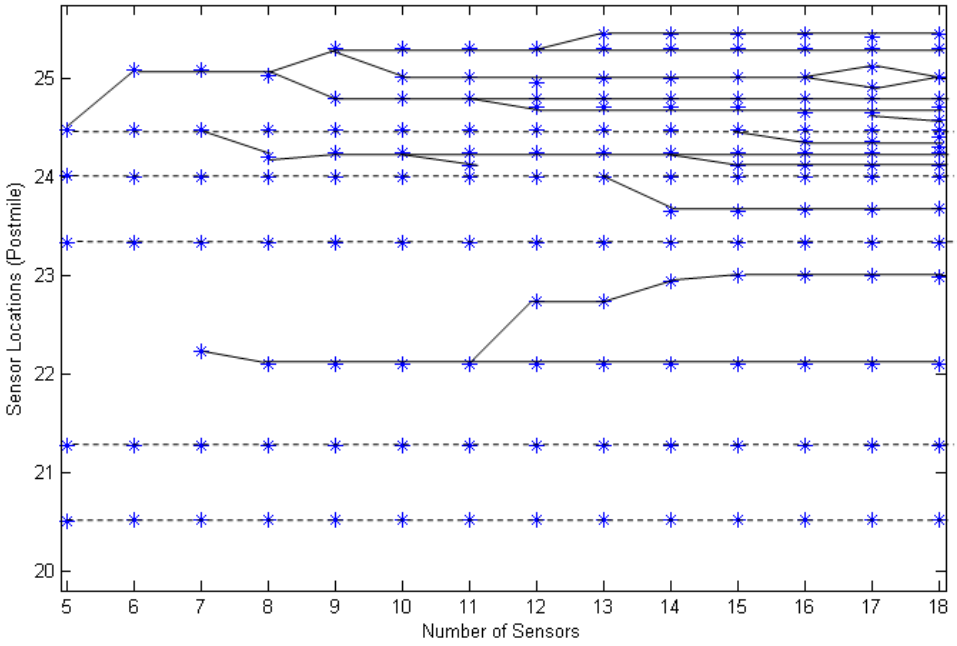


Figure 7: Evolution of Optimal Sensor Locations (from 5 to 18 Sensors)

is evident from the comparison that the benefit of using the proposed framework is significant.

6 Conclusion

We proposed a modeling framework to study optimal traffic sensor placement. The proposed method aims to capture the fact that traffic sensor deployment in reality is a continuous, evolving, and sequential process, i.e. the placement of additional sensors for new applications has to consider sensors that have already been deployed. In this article, we adopted the Dynamic Programming (DP) modeling framework recently developed by the authors for optimal sensor deployment, which has the capability to optimally deploy additional sensors by considering existing sensors. For two traffic applications: ramp metering and travel time estimation, we showed, using the Mobile Century GPS data on a real-world freeway route, that the proposed framework can generate optimal sensor placement for both applications in a sequential yet coherent manner. The generated optimal sensor placement matches well with the congestion areas of the study route, which further shows that the placement is reasonable. The optimal placement generated by the model is also superior to existing sensor placement in terms of providing better estimates to freeway travel times.

The proposed method and algorithm need to be further tested and validated on more and larger real-world traffic applications. Research in this direction will be pursued in future work and results will be reported in subsequent articles.

References

[1] H. Yang and J. Zhou. Optimal traffic ccounting locations for origin-destination matrix estimation. *Transportation Research B*, 32(1):109–126, 1998.

[2] L. Bianco, G. Confessore, and P. Reverberi. A network based model for traffic sensor location with implications on O-D matrix estimates. *Transportation Science*, 35(1):50–60, 2001.

[3] K. Ozbay, B. Bartin, and S. Chien. South Jersey real-time motorist information systems: Technology and practice. *Transportation Research Record*, 1886:68–75, 2004.

[4] I. Fujito, R. Margiotta, W. Huang, and W.A. Perez. The effect of sensor spacing on performance measure calculations. In *Proceedings of the 85th Annual Meeting of Transportation Research Board (CD-ROM)*, 2006.

[5] J. Kwon, B. McCullough, K. Petty, and P. Varaiya. Evaluation of PeMS to improve the congestion monitoring program. Technical report, Final Report for PATH TO 5319, 2006.

[6] X. Ban, Y. Li, A. Skabardonis, and J.D. Margulici. Performance evaluation of travel time methods for real time traffic applications. In *Proceedings of the 11th World Congress on Transport Research (CD-ROM)*, 2007.

[7] G. Thomas. The relationship between detector location and travel characteristics on arterial streets. *Institute of Transportation Engineers Journal*, 69(10):36–42, 1999.

[8] S. Oh, B. Ran, and K. Choi. Optimal detector location for estimating link travel time speed in urban arterial roads. In *Proceedings of the 82nd Annual Meetings of the Transportation Research Board (CD-ROM)*, 2003.

[9] S. Jung, A. Toppen, and K. Wunderlich. The effect of average loop detector spacing on the accuracy of calculated travel times: Twin cities case study. Technical report, Mitretek Systems,, 2005.

[10] H.D. Sherali, J. Desai, and H. Rakha. A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. *Transportation Research B*, 40:857–871, 2006.

[11] B. Bartin, K. Ozbay, and C. Iyigun. A clustering based methodology for determining the optimal roadway configuration of detectors for travel time estimation. *Transportation Research Record*, 2000:98–105, 2007.

[12] X. Ban, R. Herring, JD Margulici, and A. Bayen. Optimal sensor placement for freeway travel time estimation. *Submitted to the 18th International Symposium on Transportation and Traffic Theory*, 2008.

[13] S.M. Eisenman, X. Fei, X. Zhou, and H.S. Mahmassani. Number and location of sensors for real-time network traffic estimation and prediction: A sensitivity analysis. *Transportation Research Record*, 1981:253–259, 2006.

[14] NET. System wide adaptive ramp metering - high level design. Technical Report Final Draft, Prepared by NET for Caltrans and FHWA, 1996.

[15] E. Smaragdis and M. Papageorgiou. Series of new local ramp metering strategies. *Journal of the Transportation Research Board*, 1856:74–86, 2003.

[16] L. Chu, X. Liu, W. Recker, and H.M. Zhang. Performance evaluating of adaptive ramp metering algorithms using microscopic traffic simulation model. *Journal of Transportation Engineering*, 130(3):330–338, 2004.

[17] S. Ahn, R.L. Bertini, B. Auffray, J.H. Ross, and O. Eshel. Evaluating the benefits of a system-wide adaptive ramp-metering strategy in portland. *Journal of Transportation Research Board: Transportation Research Record*, 2012:47–56, 2007.

[18] S. Amin et al. Mobile century-using gps mobile phones as traffic sensors: a field experiment. In *Proceedings of the 15th World congress on ITS*, 2008.

[19] D. Work, O.P. Tossavainen, S. Blandin, A. Bayen, T. Iwuchukwu, and K. Tracton. An ensemble kalman filtering approach to highway traffic estimation using gps enabled mobile devices. In *Proceedings of the 47th IEEE Conference on Decision and Control*, 2008.

[20] http://pems.eecs.berkeley.edu/public/.

An Integer Linear Program for Optimizing Sensor Locations along Corridors

Henry X. Liu
Department of Civil Engineering
500 Pillsbury Drive S.E.
Minneapolis, MN 55455
Phone: (612) 625-6347
Email: henryliu@umn.edu

Adam Danczyk
Department of Civil Engineering
500 Pillsbury Drive S.E.
Minneapolis, MN 55455
Phone: (612) 625-0249
Email: danc0010@umn.edu

Submitted to Transportation Research, Part B

June 1, 2008

ABSTRACT

Accurate and reliable measurements of system states are critical for intelligent operations. However, budgetary restrictions often forbid the sensor instrumentation that provides complete and exhaustive measurements of a studied system. The purpose of this paper is to address this problem in the context of a transportation freeway corridor, answering the question of how to allocate sensor resources with a limited budget. The work here focuses on optimally deploying aggregated-point sensors, such as inductance loop detectors, on a one-directional freeway corridor to measure the traffic condition and identify freeway bottlenecks. To date, only a nonlinear model with integer variables exists to solve this type of problem. This paper transforms this model into an equivalent integer linear program, which can be solved using resource constrained shortest path algorithms. In addition, a secondary objective is incorporated into this model to account for sensor failures and a multiobjective program is formulated to help identify feasible designs that find balanced tradeoffs. This developed model can serve as a planning tool for sensor installments on corridors with or without existing sensor configurations.

INTRODUCTION

Accurate and reliable measurements of system states are critical for intelligent operations. The reliability and accuracy of this data is greatly dependent on the allocation of sensors throughout the system to measure on-site conditions. Generally, as the density of sensors increases, the overall accuracy of an assessed performance measure tends to improve. In an ideal, monetary-free world, practitioners would instrument a maximal number of sensors to find the ground-truth state of a system. Unfortunately, budgetary constraints often forbid such lavish instrumentation. Practitioners are instead left to seek out the most optimal placement of their sensors, in terms of measuring the overall system's state accurately given their limitations. However, finding this optimal placement in a highly complex environment is not often an easy achievement.

The purpose of this research is to develop a model for optimally allocating sensors along a one-directional corridor for accurate performance monitoring purposes. One-directional corridors represent a wide variety of real world environments, including transportation freeways, hydrologic river systems, and telecommunications traffic. The focus of this work is exclusively on aggregated-point sensors, a type of sensor that has been historically more affordable to deploy. These sensors, which include inductance loop detectors, have inherent limitations to the extent of data that is collectable and, as a result, most practitioners have adopted a rule of thumb to only assess conditions between two neighboring sensors. Earlier research has developed a nonlinear model to seek an optimal configuration given these limitations, but the model's non-convex objective function makes it difficult to solve with traditional solver. This paper overcomes this issue by transforming the problem into a linearized model, allowing a solution to be found with greater ease through traditional solvers.

The sensor location problem is a topic that has been rigorously studied in the past for a wide variety of applications. Most literature focuses exclusively on transportation problems. Ozbay et al. (2004) studied the quality of travel time estimation when compared with sensor locations under recurrent and non-recurrent congestion. Kwon et al. (2006) created an empirical model that relates roadway-based sensor spacing to the accuracy of measuring traffic congestion by studying that overall accuracy falters as the distance between sensors increases. Ban et al. (2007) showed similar results, illustrating that increasing the sensor spacing causes higher travel time estimation errors and higher variations in travel time reliability. Bartin et al. (2006) also focused on roadway sensor spacing, finding that the marginal gain of travel time accuracy decreased as the number of road-based surveillance units increased.

A study by Fujito et al. (2006) determined that the actual location of sensors is more important for the estimation of congestion along a transportation corridor than uniform spacing. Empirical analysis showed that results varied accordingly to the positions in which sensors were removed. This strategic location concept can be seen in numerous transportation sensor location problems. Several literatures develop models for determining instrumentation location on transportation networks for identifying the most origin-destination (O-D) paths (Fei and Mahmassani, 2007, Fei et al., 2007, Bianco et al.,

2001, Yang and Zhou, 1998, Yang et al., 2006). Others focus on sensor allocation on a transportation network for toll collection purposes (Zhang and Yang, 2004) and for traffic monitoring to reduce network risks (Gendreau et al., 2000), such as policing for drunk drivers.

Sherali et al. (2006) developed a model for optimally allocating Automatic Vehicle Identification (AVI) tag readers along a transportation corridor. AVI tag readers are a type of reidentification sensor that can track an individual vehicle as it moves from sensor to sensor. This research assumed that an environmental characteristic, called benefit, exists between any two sites and can be captured by allocated sensors at these sites. In this case, the environmental characteristic was deemed to be travel time variability. The model has a quadratic objective and linear constraints, encompassing budgetary restrictions. This research designed an optimization approach based on a Reformulation-Linearization technique combined with semi-definite programming concepts that formulate the problem in linearized form. This formulation is similar to that found in (Adams and Sherali, 1986, Watters, 1967). It effectively gives a linearized approach to solving the sensor location with reidentification sensors.

Liu and Danczyk (2007) conducted similar research on a one-directional transportation corridor, except using inductance loop detectors instead of AVI tag readers. Loop detectors are aggregated-point sensors, meaning they only report back aggregated conditions at their site and seldom have any ability to track an individual vehicle between sites. It was also assumed that an environmental characteristic, called benefit, existed between any two sites. A ‘Neighboring Sensor’ assumption was made, meaning that benefit could only be captured between sensors that were neighbors, to keep in line with the state of the practice. This assumption creates a dynamic coefficient for benefit that, when coupled with the intuitive quadratic objective, creates a nonlinear program which unfortunately cannot use previous Reformulation-Linearization techniques to linearize the problem. Consequently, a heuristic must be used.

This paper builds upon previous work done by Liu and Danczyk (2007) by taking the formulated nonlinear model and seeking methods to more efficiently and accurately find a solution through linearization. It is intended to bridge the gap between other research conducted on the sensor location problem by tying an existing aggregated-point sensor location model to a simplified, linearized case. It will start by discussing the existing nonlinear model. From there, it will develop a linearized version of the model that produces an identical solution. Lastly, it will propose a means to account for sensor failures in the model and define a multiobjective approach for finding a good medium.

NONLINEAR MODEL

The purpose of this research is to allocate aggregated-point sensors along a one-directional corridor to optimize performance measure accuracy. This paper will focus on a specific problem—allocating loop detectors along a transportation freeway corridor for freeway bottleneck identification and assessment. The corridor considered is a typical one, such as the one shown in Figure 1. The freeway receives an entry flow rate at its

upstream point (q_0) during the study period, T. Along the route, entrance ramps add additional flow (q_1, q_3, q_5) while exit ramps reduce flow (q_2, q_4, q_6). The freeway is divided into cells not necessarily of equal length. Each cell i is designated as a potential site to place a sensor. Generally speaking, this sensor would be located in the middle of the cell. It is assumed that between any two cells is a constant, predetermined environmental characteristic, or benefit (b_{ij}).

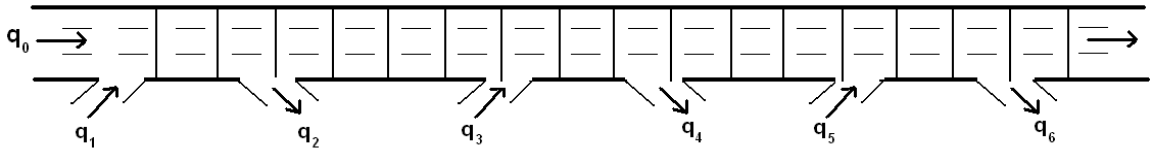


Figure 1: A freeway sketch, divided into cells

The model is shown in Section (1).

$$\textit{Maximize}: \quad \sum_{i \in A} \sum_{j \in A} b_{ij} y_i y_j \tag{1a}$$

$$\textit{Subject To}: \quad \sum_{j \in N} y_j \leq R \tag{1b}$$

$$\sum_{j \in N} C_j y_j \leq B \tag{1c}$$

$$y \text{ binary} \tag{1d}$$

The objective function (1a) seeks to maximize the overall benefit acquired by a given configuration. The binary variable, y_j, is a value of 1 if a sensor is placed in cell j and 0 if not. Constraint (1b) ensures that the number of sensors placed does not exceed a maximum number of allowed sensors, R. Constraint (1c) restricts the sensors placed in any cell j with a cost, C_j, from exceeding a specified budget, B. Constraint (1d) maintains binary variable types.

This model is very intuitive. Where the difficulty arises is in its quadratic nature and exclusive classification of benefit. To uphold the Neighboring Sensor assumption that comes with aggregated-point sensors, the benefit needs to be classified as only acquirable between sites that do not have sensors nestled in the middle. Since the corridor is one-directional, benefit can only be gained in the traveling direction. The benefit, defined as an ability to detect and measure freeway bottlenecks, is formulated in Equation (2). On an eligible link, it is a constant, non-negative value that represents the average speed gradient during the study period, T. This speed gradient, which is the difference between time-dependent velocities at site i (V_i^t) and site j (V_j^t) at time t, can be found either through simulation or from historic conditions. If a sensor is present between site i and site j or the stationing location of site j (S_j) precedes the stationing location of site i (S_i), the benefit is defaulted to a value of zero.

$$b_{ij} = \begin{cases} 0 & \sum_{k=i}^{j} y_k > 2 \quad or \quad S_j \le S_i \\ \max\left\{0, \dfrac{\sum_{t=0}^{T}(V_j^t - V_i^t)}{(S_j - S_i)}\right\} & otherwise \end{cases} \qquad (2)$$

This formulation causes the benefit factor to be dynamic. That is, its value is determined by a variable rather than being constant. It essentially places the product of three variables in the objective function, which makes it difficult to solve and even more difficult to linearize. A genetic algorithm was previously used to find a solution.

Doing a direct conversion from a nonlinear problem to a linear problem with this formulation may be too difficult, if not impossible. A better method would be to rethink the original problem and reformulate with different linear components that answer the same question.

FORMULATION OF LINEARIZED SENSOR LOCATION PROBLEM

Linearization of the sensor location problem using aggregated-point sensors requires that the problem be disaffiliated from simply assigning sensors to given sites. Rather than focusing on the individual sites as eligible variables, the objective function should be based on accruing benefit gained by covering the links themselves. To cover a link (i, j), sensors must be located at site i and site j, as proposed in (Watters, 1967). For example, instead of allocating sensors to points i=2 and j=4, the covered link would be (i, j) = (2, 4). The conversion from point-based sensor coverage (using y_i and y_j) to link-based coverage (using x_{ij}) can be found in Equation (3).

$$y_i y_j \ge x_{ij} \qquad (3)$$

Equation (3) changes the quadratic portion of this model into a single variable. It implies that if a sensor is located at both site i and site j (meaning $y_i = y_j = 1$), then link (i, j) would be considered actively covered ($x_{ij} = 1$). Likewise, if either site i or site j lack sensors, then link (i, j) would be considered empty of coverage ($x_{ij} = 0$). However, just because sensors are present at site i and site j does not mean link coverage is guaranteed, as there might be a sensor at site k (i<k<j) that would violate the Neighboring Sensor assumption by having a link pass over a sensor without making a connection. This translation allows x_{ij} to become the new decision variable for the new model.

From this translation, a new objective value is produced, using a single variable rather than two, and a new constraint for that variable.

$$Maximize: \quad \sum_{i \in A} \sum_{j \in A} b_{ij} x_{ij} \tag{4a}$$

$$x \ binary \tag{4b}$$

Equation (4a) takes the first step in linearizing the model, maximizing the benefit gained from covering links instead of individual points. However, this does not fix any of the problems. By using link coverage instead of point coverage and offering no translation between the two, the existing constraints—(1b) and (1c)—cannot bound the problem anymore. Also, there is no guarantee that x_{ij} will follow the Neighboring Sensor assumption by itself, as there are no constraints to bound it accordingly. As a result, the dynamic nature of the benefit value, b_{ij}, must be kept in the model to constrain the objective function from offering more benefit than is truly available. To deal with these issues, new intuitive constraints must be formulated.

The Neighboring Sensor assumption states that only neighboring sensors can be linked together. To exemplify this point, consider a line of 20 sensors arranged sequentially in 20 cells. With the stated assumption, the only links connected to the 10^{th} sensor would be link (9, 10) and link (10, 11), as any other links would cross over a sensor. It can be seen that, at the 10^{th} sensor, one link at most enters that site from an upstream sensor while one link at most leaves that site for a downstream sensor. This is true for all sensors, aside from the first and last one for now. This formulates Equations (5a) and (5b).

$$\sum_{i=i'}^{N} x_{ij} \leq 1 \quad \forall j \tag{5a}$$

$$\sum_{j=1}^{j'} x_{ij} \leq 1 \quad \forall i \tag{5b}$$

Similarly, it is known that links can only originate from or be destined to sites that have allocated sensors. With at least one link sprouting from such a site, a relationship can be made between sensor presence and link coverage, as shown in Equation (5c). This sensor presence will allow the original constraints, (1b) and (1c), to be included as is in the new formulation.

$$\sum_{j=i+1}^{j'} x_{ij} - y_i = 0 \quad i \in 1,..., N \tag{5c}$$

As is, the problem is open ended. Without definite boundaries, it could easily go on forever. The locations of the first and last sensor along the corridor are unknown, so boundaries cannot be statically established based on them. Instead, an assumption will be made that a sensor—a phantom, nonexistent sensor—is present at both site i' and site j', where site i' is a site existing immediately before the start of the analyzed corridor and site j' is a site existing just after the terminus of the analyzed corridor. These locations are shown in Equations (6a) and (6b).

$$i' = 0 \tag{6a}$$

$$j' = N + 1 \tag{6b}$$

If looking at these sites from a point-based assignment perspective, like for the nonlinear model, both values would be non-zero to represent the presence of a sensor. Equations (7a) and (7b) illustrate this point.

$$y_{i'} = 1 \tag{7a}$$

$$y_{j'} = 1 \tag{7b}$$

However, the focus of this research is not based on the point-based coverage, but rather the link-based coverage. To redefine Equations (7a) and (7b), their meaning needs to be thought of in terms of link-based coverage. These constraints imply that at least one link originates or terminates at either site, even without any other sensors placed on the network. This translates to Equations (8a) and (8b).

$$\sum_{j=1}^{j'} x_{i'j} = 1 \tag{8a}$$

$$\sum_{i=i'}^{N} x_{ij'} = 1 \tag{8b}$$

Intuitively, at any real sensor site, one link must be inbound to that site and one must be outbound. In the earlier example, at the 10^{th} sensor, one link was coming from the 9^{th} sensor and one link was going to the 11^{th} sensor. Since all real sensors now have a neighboring sensor, either real or phantom, both upstream and downstream, this holds true for all real sensors.

$$\sum_{i=i'}^{j} x_{ij} = \sum_{k=j}^{j'} x_{jk} \quad \forall\ j \in 1,\ldots,N \tag{9}$$

Equation (8a) connects one link from a phantom sensor to some other real or phantom sensor. Equation (9) ensures that any real sensor receiving a link from upstream produces another link heading downstream. Equation (8b) connects one link from the previous real or phantom sensor to the last phantom sensor. Since only one active link originates from the first phantom sensor, the maximum number of links passing over or connecting to any site becomes a value of one at most by default. With that, Equations (5a) and (5b) become redundant and can be dropped from the formulation. Additionally, since the opportunity for links to cross over sensors has been removed, the variable portion of the benefit factor can be removed, since other constraints prevent a violation from occurring. The benefit factor can be redefined, as seen in Equation (10). Since it is determined with only constants, the benefit factor becomes a constant value.

$$b_{ij} = \begin{Bmatrix} 0 & S_j \le S_i \\ \max\left\{0, \quad \dfrac{\sum_{t=0}^{T}(V_j^t - V_i^t)}{(S_j - S_i)}\right\} & otherwise \end{Bmatrix} \tag{10}$$

Since this problem is one-directional, meaning that flow runs from upstream to downstream only, certain link-based variables can be immediately disregarded from the general group as offering no possible coverage. Equation (11) sets all link-based variables to a value of zero if their destination precedes or is their origin. It reduces the number of decision variables from n^2 to $n \times (n-1)/2$, assuming there are n eligible locations for sensor placement. The final solution would be the same with this constraint, as no benefit is gained even if these links are covered.

$$x_{ij} = 0 \quad \forall \{i \ge j\} \tag{11}$$

With all of this combined, the sensor location problem for aggregated-point sensors is linearized, as shown below in Section (12).

$$Maximize: \quad \sum_{i \in N}\sum_{j \in N} b_{ij} x_{ij} \tag{12a}$$

$$Subject\ To: \quad \sum_{j=i+1}^{j'} x_{ij} - y_i = 0 \quad i \in 1,...,N \tag{12b}$$

$$\sum_{i=i'}^{j} x_{ij} = \sum_{k=j}^{j'} x_{jk} \quad \forall\ j \in 1,...,N \tag{12c}$$

$$\sum_{j=1}^{j'} x_{i'j} = 1 \tag{12d}$$

$$\sum_{i=i'}^{N} x_{ij'} = 1 \tag{12e}$$

$$\sum_{j=1}^{N} y_j \le R \tag{12f}$$

$$\sum_{j=1}^{N} C_j y_j \le B \tag{12g}$$

$$x_{ij} = 0 \quad \forall\ i \ge j \tag{12h}$$

$$x\ binary \tag{12i}$$

Although this new formulation deals with more constraints and variables than its nonlinear counterpart, it is much easier to solve. In fact, a means to solve this problem through the use of a resource constrained shortest path algorithm exists and is formulated in the next section.

SOLUTION ALGORITHM

One way to solve this type of problem would be to use a traditional solver. Another way, however, would be to represent this problem graphically and solve using a constrained shortest path search algorithm, since the formulation of this problem is mathematically similar to the formulation of shortest path problems. Considering a typical corridor, a graphical representation of potential sensor allocations can be easily generated to represent this problem. This graph is denoted as G(N, A), with a node set N (corresponding to the variable y) and a directed arc set $(i, j) \in A$ (corresponding to the variable x). Since the problem only focuses on a single direction, the only available arcs run between sites that follow that direction. In other words, an arc exists between site i and site j, where site i precedes site j, but not the other way around.

The end result is a network of paths running between the two phantom sensor sites. Between the two phantom sensors, the total number of paths represents all the enumerable configurations. An example of this can be seen in Figure 2, where three real sites exist on a certain freeway network.

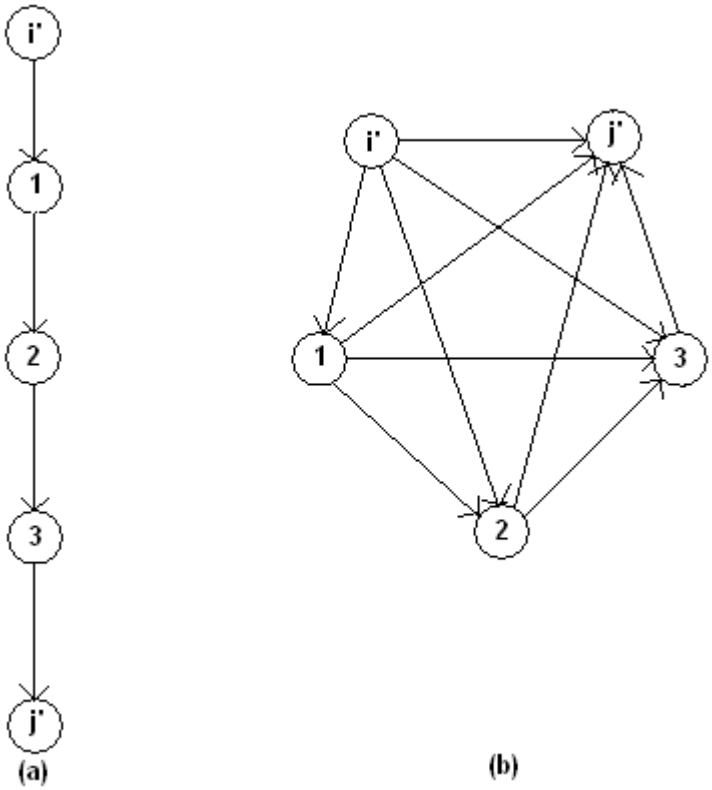


Figure 2: Illustration of a Freeway Corridor (a) and its corresponding graph G(N, A) (b)

A shortest path problem minimizes the cost of using a path between a given origin and destination, given that one unit of flow leaves the source node, one unit of flow arrives at

the sink node, and all flows entering transverse nodes also depart that node. For the case of the linearized model, the objective function maximizes benefit captured, which equals the minimization of cost, or negative benefit. Constraints (12d) and (12e) can be considered source and sink nodes, respectively, while constraint (12c) can represent the balance at any transverse node.

Many algorithms to find shortest paths have been researched in the past (Denardo and Fox, 1979, Dial, 1969, Pape, 1974). One of the more well-known is the Bellman-Ford-Moore algorithm (Bellman, 1958, Ford and Fulkerson, 1958, Moore, 1959), which is a one-to-all search procedure. Another well-known one is Dijkstra's algorithm (Dijkstra, 1959), which is much more efficient for one-to-all cases (Rardin, 1998). These algorithms operate without any form of resource constraint.

If a shortest path algorithm was used to find an optimal sensor configuration and that path was shown to be within the allotted budgetary constraints (12f) and (12g), then that solution would be feasible and optimal. However, if a constraint is violated and the cause of the violation cannot be reasonably excluded from the problem, then the problem becomes a resource constrained shortest path (RCSP) problem, with Equations (12f) and (12g) serving as additional constraints to a standard shortest path problem. RCSP problems are defined as NP-Hard problems and, consequently, much more difficult to solve. Numerous algorithms have been proposed for solving RCSP problems (Witzgall and Goldman, 1965, Avella et al., 2002, Handler and Zang, 1980, Joksch, 1966, Jensen and Berry, 1971). Aneja et al. (1983) developed an algorithm based on reduction tests and a generalization of the Dijkstra algorithm. Beasley and Christofides (1989) developed a Branch-and-Bound approach based on a lagrangean heuristic for the solution of RSCP with more than one resource constraint.

There are several means to produce an optimal solution for this problem. Now, the question is whether this optimal solution would truly be the best in an imperfect world. For example, the mechanical failure of a sensor would detrimentally impact the accurate measuring abilities of this configuration. With this in mind, how can the model be modified to account for faulty sensors?

ADJUSTMENT FOR FAILING SENSORS

Failing sensors can be a serious problem when attempting to measure the state of a corridor. Unfortunately, it is far from uncommon. For example, a study conducted by the State of New York found that, of the 15,000 existing roadway inductive loop detectors maintained by the state, 25 percent were not operating at any given time (Traffic Detector Handbook: Third Edition, 2006). This failure rate is consistent with other failure literature.

Losing a sensor can detrimentally impact the accuracy of measurements. Moreover, the loss of certain sensors, placed in critical locations, can cause an even greater impact on performance measure accuracy. When an aggregated-point sensor goes offline, generally the two adjacent aggregated-point sensors make an estimate of conditions present in the

region where the failed sensor once operated. However, if those adjacent sensors are very far apart, the accuracy can greatly diminish. If the goal is to guarantee a minimal impact of a failed sensor, it would be much more beneficial to allocate sensors in a uniformly-spaced configuration. Doing this could contradict the strategic placement aspect that the sensor location problem aims to find in a non-uniform environment. Thus, there may be two competing goals that need to be addressed.

Solving this problem requires that some basic assumptions are known. First, it is assumed that a sensor failure is something that can be identified. That is, a failure does not occur and not get accounted for. Second, it is assumed that the probability of successful data collection for a sensor type d, P_d, is known. Third, the probability of successful data collection at a given site i, $P_{site(i)}$, is known. With this knowledge, the probability of acquiring successful data from a given site with a given sensor type, $P_{s(i)}$, can be computed, which is shown in Equation (13a). Since this is computed with constant values, it is a constant value. For example, if a given sensor has a successful operation probability of 90 percent and a given site i is known to have a successful operation probability of 80 percent, then the probability that the sensor would produce functional data would be 72 percent.

$$P_{s(i)} = P_d * P_{site(i)} \tag{13a}$$

It will be assumed that when a sensor fails at k, sensors at any site i and site j can collect data in its place. This assumes i<k<j. A probability of this being the case can be found by finding a value of $P_{S_{ijk}}$, which is shown in Equation (13b). For example, if sites i, j, and k were known to have successful operation probabilities of 95 percent each, then the probability that site k fails and sites i and j would remain operational to capture data would be 4.5 percent.

$$P_{S_{ijk}} = P_{S(i)} * P_{S(j)} * (1 - P_{S(k)}) \tag{13b}$$

Next, the impact of a failed sensor must be defined. Conceptually, a sensor failure would cause a loss of accurate information, or a reduction in benefit. However, if other sensors are on the corridor, they can potentially pick up some of the lost accuracy, or regain a fraction of the lost benefit. In a configuration set up to account for intermittent sensor information losses, the gap between the benefit lost and the benefit recovered would be small. This gap is defined in Equation (13c).

$$Impact\ Value = b_{ik} + b_{kj} - b_{ij} \quad (i<k<j) \tag{13c}$$

This impact would only matter if other sensors were available at site i and site j to pick up the lost information from site k. Therefore, this impact factor must be multiplied by a variable, w_{ijk}, which represents the presence of sensors at all three sites. It is based on several criteria to ensure that it truly represents situations where the necessary sensors are present. These criteria are outlined in Section (14).

$$w_{ijk} \leq y_i \ \ \forall i,j,k \quad (14a)$$
$$w_{ijk} \leq y_j \ \ \forall i,j,k \quad (14b)$$
$$w_{ijk} \leq y_k \ \ \forall i,j,k \quad (14c)$$
$$w_{ijk} \geq 0 \ \ \forall i,j,k \quad (14d)$$
$$y_i + y_j + y_k - w_{ijk} \leq 2 \ \forall i < k < j \quad (14e)$$

As is, the impact of a failed sensor can be found by multiplying its marginal impact by the presence of adjacent sensors to capture the lost information and then by multiplying further by the probability of such an incident occurring. However, this could be an underestimation of the true potential impact. This assumes that the immediately adjacent sensors are functional, which may not always be true. In fact, those adjacent sensors may have failed and a new pair of sensors beyond the original two must be sought after. So, not only should impact be assessed based on the probability of a failure with functional neighboring sensors, but also the probability of those neighboring sensors failing to offer support. The probability of failures along link (i, j) given a failed k, $P_{F_{\{ijk\}}}$, is described in Equation (15a).

$$P_{F_{\{ijk\}}} = \prod_{(l,m)}\left(1 - P_{S_{lmk}} w_{lmk}\right) \quad \begin{cases} I = \{i, \dots, k-1\} \\ J = \{k+1, \dots, j\} \\ (l,m) \in I \times J \\ (l,m) \neq (i,j) \end{cases} \quad (15a)$$

The entire objective can be formulated to define the absolute benefit loss because of sensor failures across the corridor. It is shown in Equation (15b).

$$\sum_{k=1}^{N}[\sum_{i=i'}^{k-1}\sum_{j=k+1}^{j'} P_{S_{ijk}} * P_{F_{\{ijk\}}} * [b_{ik} + b_{kj} - b_{ij}] * w_{ijk}] \quad (15b)$$

There is a clear problem—this is a nonlinear formulation. Being so, it potentially takes away the benefits gained by linearizing the original problem. Unfortunately, there does not appear to be a means to linearize this portion. The only way to get around it is to make an assumption that only one sensor can fail at a time, thus removing the need to assess whether two sensors fail in sequence and the means to resolve that issue. With that, the failure probability can be dropped from the general formula. The new objective function is shown in Equation (16a) instead of Equation (15b).

$$\sum_{k=1}^{N}[\sum_{i=i'}^{k-1}\sum_{j=k+1}^{j'} P_{S_{ijk}} * [b_{ik} + b_{kj} - b_{ij}] * w_{ijk}] \quad (16a)$$

Similarly, to ensure that site i and site j are only adjacent sensors, Equation (14e) must be replaced with Equation (16b) while keeping Equations (14a), (14b), (14c), and (14d) in the constraints.

$$y_i + y_j + y_k - w_{ijk} - \sum_{\mathrm{l=i+1}}^{\mathrm{j-1}} \mathrm{y_l} + y_k \leq 2 \quad \forall i < k < j \quad (16b)$$

Clearly, the goal of this formulation would be to minimize the overall benefit loss caused by sensor failure. This would be occurring while the original objective is maximizing the benefit gained. Thus, a multiobjective program is required to help find an appropriate balance between the two. However, to do this effectively, one must know how and where to find an appropriate and efficient balance between the two.

One of the popular means for selecting a balance of conditions is to identify the Pareto Set, or Pareto Frontier. Given certain corridor conditions, allocating sensor resources to make one goal better without making the other goal worse is called a Pareto improvement. When no further Pareto improvements can be made, the allocation is called Pareto efficient. The set of choices that are all Pareto efficient make up the Pareto Frontier. Research has been conducted in the field of approximating these Pareto efficient points (Wilson et al., 2000, Kasprazak and Lewis, 1999, Li et al., 1998, Das and Dennis, 1997, Das and Dennis, 1998).

For this sensor location problem, two objectives are in competition with one another. The first maximizes overall benefit based on a strategic sensor configuration. The second minimizes the benefit lost by failed sensors. These two objectives, shown in Equations (17a) and (17b), use constraints (12b-12j), (14a-14d), and (16b).

$$Maximize: \quad \sum_{i \in N} \sum_{j \in N} b_{ij} x_{ij} \tag{17a}$$

$$Minimize: \quad \sum_{k=1}^{N} [\sum_{i=i'}^{k-1} \sum_{j=k+1}^{j'} P_{S_{ijk}} * [b_{ik} + b_{kj} - b_{ij}] * w_{ijk}] \tag{17b}$$

CASE STUDY

A case study will be conducted on a real transportation corridor to illustrate how the model allocates sensor resources for the different objectives. In this case, loop detectors are being allocated along Interstate 94 (I-94) in Minneapolis, Minnesota, for freeway bottleneck detection purposes. This corridor is 7.2 miles in length. Benefit is predetermined as the success rate and timeliness of sensing an active bottleneck between any two sites, shown in Equation (10). Conditions on this road were simulated using the cell transmission model developed by Daganzo (1994) and Daganzo (1995) with real data captured from entry and exit flows measured on that corridor on Wednesday, June 13, 2007, during the PM peak rush hour. These loop detectors are assumed to have a constant successful operation probability of 95 percent for all sites. 16 loop detectors are currently implemented along this corridor, as seen in Figure 3.

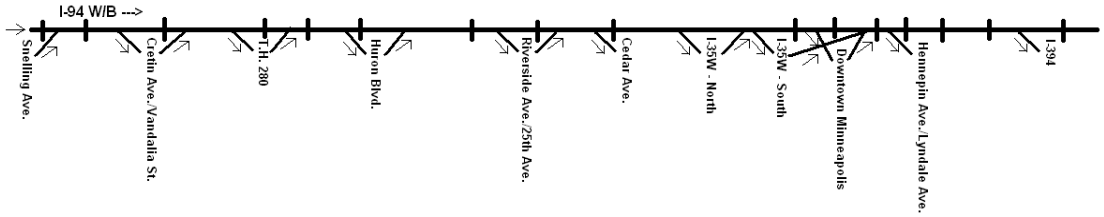


Figure 3: Existing loop detector configuration on westbound I-94. Loop detectors are represented by vertical segments.

It will be assumed that all of these loop detectors have been removed and new ones will be implemented, based on recommendations by the model. 16 new loop detectors are available for placement. The model allocates them as shown in Figure 4.

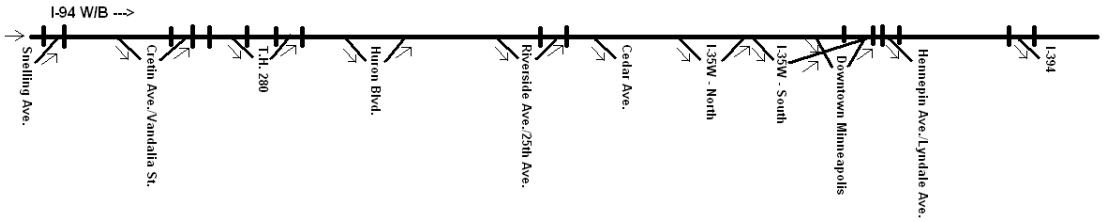


Figure 4: New loop detector configuration, based on the model's recommendations for 16 available loop detectors. Loop detectors are represented by vertical segments.

The loop detector configuration is much more clustered with the model's recommendations than the original installation. This is as expected, since, for bottleneck identification, sensors are better when allocated close the bottleneck site. The overall benefit also increases significantly, jumping from 86.8 to 1085.5 between the existing and new configuration, respectively. As shown in Figure 5, the total benefit gained increases as the number of sensors, configured to the optimal configuration, increases. The marginal rate of return decreases with each new sensor, as the strongest bottlenecks are covered first because they offer the highest benefit.

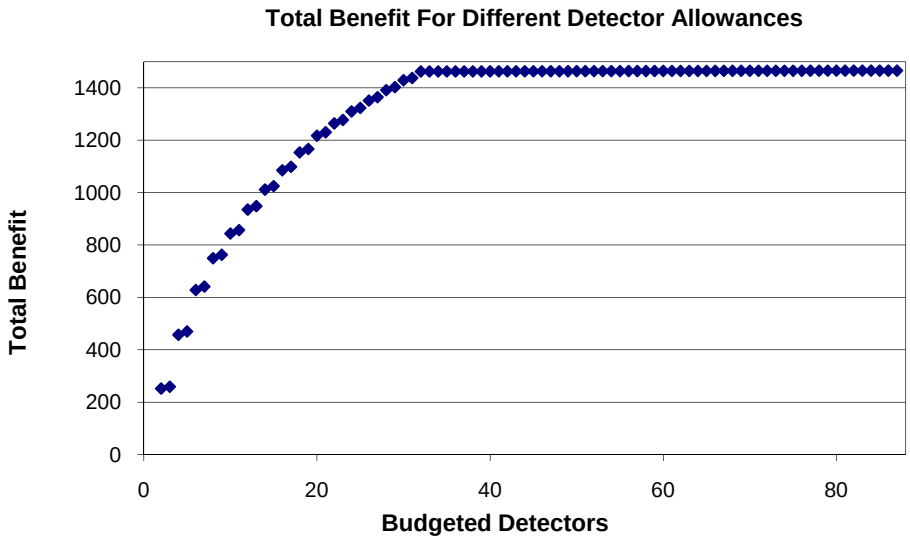


Figure 5: Total benefit gained for optimal allocations with various sensor budgets

More often, new sensors are added to corridors with an existing sensor configuration. Given that the existing sensors cannot be moved, the model can allocate new sensors around this static configuration to optimize benefit despite the constraints. 16 new loop detectors were added incrementally to the I-94 network and their total resulting benefit is shown in Figure 6. When the first new sensor is added to the existing configuration, the overall benefit increases substantially. Afterwards, each new sensor brings incremental overall benefit that decreases marginally.

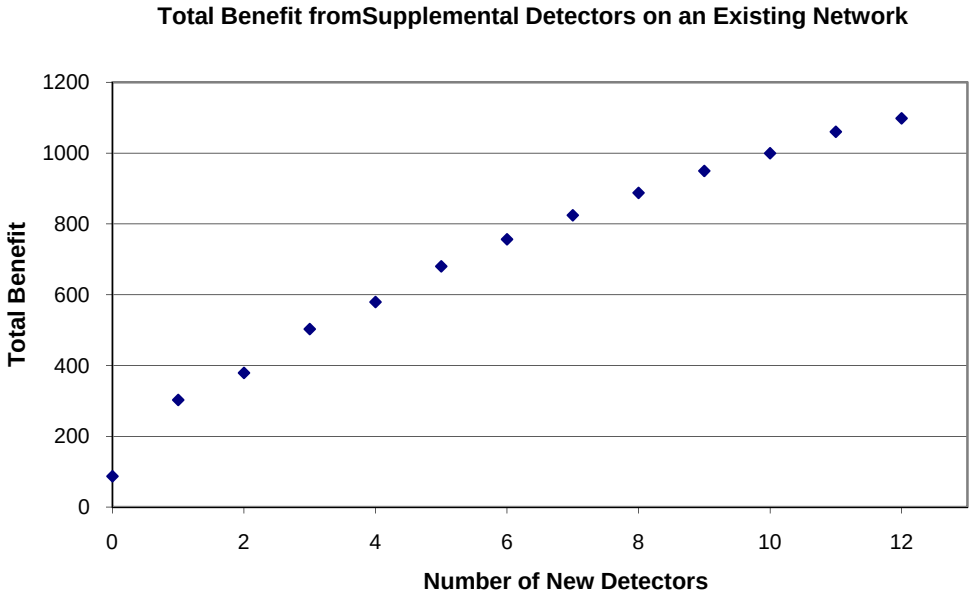


Figure 6: Total benefit gained by adding new sensors to the existing I-94 configuration

The resulting benefit captured, for either an empty corridor or one with an existing configuration, has been illustrated in Figure 5 and Figure 6, but this shows nothing of how sensor allocation would be to accommodate failing sensors. Thinking of the I-94 network again, how would loop detectors be allocated to minimize benefit loss if 16 loop detectors are budgeted and the network has no previous allocated sensors? Figure 7 reflects the configuration the model recommended to reduce benefit loss. This configuration yields a total benefit of 32.5, which is only 3 percent optimal when compared with the best configuration for the same number of sensors.

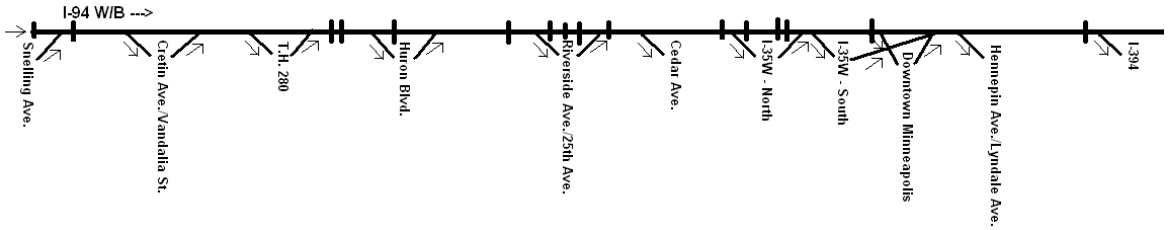


Figure 7: Allocation of loop detectors to minimize benefit loss due to sensor failures

As seen in Figure 4 and Figure 7, the configurations are very different when using different objectives. Thus, a multiobjective program is needed to determine configurations that are efficient tradeoffs between the two objectives. For this work, the procedure defined by Wilson et al. (2000) will be used to explore the design space and identify a rich set of potential points along the Pareto Frontier. 4,000 feasible designs will be picked at random and their solutions for both objectives will be calculated. Figure 8 shows the available designs generated for this problem and the resulting Pareto Frontier.

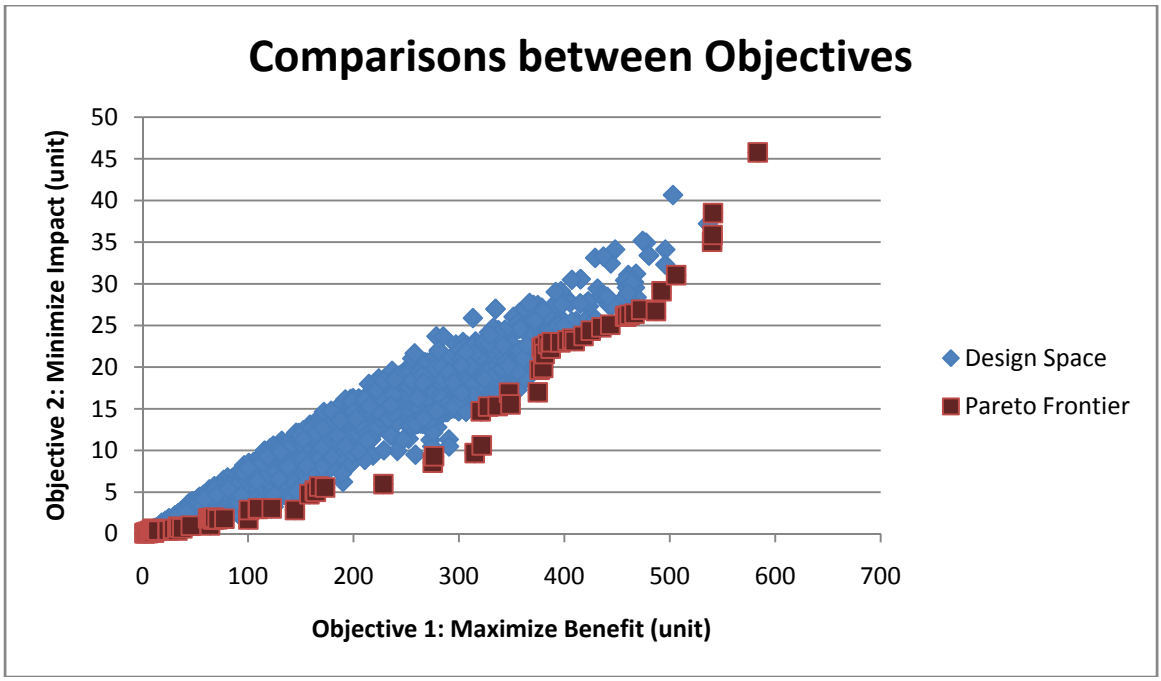


Figure 8: Comparison between Objectives

The Pareto Frontier points fall where expected. That is, they fall in a region where the first objective (17a) is maximized and the second objective (17b) is minimized. Each point represents a design where resources could not be allocated to benefit one objective without worsening the second objective. It is clear that, for this problem, designs that strongly capture strategic benefit in the first objective generally come with a high impact of failed sensors in the second objective and vice versa. Selecting the appropriate balance now requires personal preference from the engineer or planner who is designing the deployment. However, he or she has a much better set of efficient points with which to work than before.

CONCLUSION

This paper has addressed the issue of allocating sensors along a one-dimensional corridor where a predetermined environmental characteristic is known between any two sites. A linearized optimization model for aggregated-point sensors was proposed, based on a previous nonlinear model solving the same problem. This linearized model is far superior over its nonlinear counterpart, as it can use a traditional solver or a resource constrained shortest path algorithm to find the optimal solution versus using an unreliable heuristic for solving the complex, nonlinear model.

The work done in this paper has widespread applications in a variety of fields. Interested parties would strongly benefit from this linearized model because they would find the best locations for their sensors and potentially receive the most accurate assessments of their target environment, all with very little difficulty. This model can be applied to both

new and existing infrastructure projects. On an operations level, it would be assistive by answering the following questions:

- Given a corridor with a certain existing sensor configuration, where should new sensors be allocated to better measure the operational state of the system?
- When dealing with a corridor with ailing sensors present and a limited budget for replacement, which sensors should be prioritized for replacement?
- Given a new, sensor-free corridor, such as a roadway arterial transformed into a freeway link, how should sensors be allocated to best measure the system's operational state?

ACKNOWLEDGEMENTS

This paper is part of an ongoing project with the California Department of Transportation (Caltrans) to study the disparity of freeway performance measurements with varied loop detector placements surrounding freeway bottlenecks. Financial support for conducting this research was provided by Caltrans and the University of California, Berkeley. Sincere thanks go to Mr. JD Margulici and Dr. Jeff Ban of UC-Berkeley for their contributions and collaboration in this project.

REFERENCES

1.) Adams, W., Sherali, H. (1986), A Tight Linearization and an Algorithm for Zero-One Quadratic Programming Problems. Management Science, Vol. 32, No. 10, pp. 1274-1290.

2.) Aneja, Y.P., Aggarwal, V., and Nair, K.P.K (1983), Shortest chain subject to side conditions, Networks, Vol. 13, pp. 295-302.

3.) Avella, Pasquale, Boccia, Maurizio, and Sforza, Antonio (2002), A penalty function heuristic for the resource constrained shortest path problem, European Journal of Operational Research, Vol. 142, No. 2, pp. 221-230.

4.) Ban, J., Li, Y., Skabardonis, A., Margulici, J.D (2007), Performance Evaluation of Travel Time Methods for Real Time Traffic Applications. In *Proceedings of the 11th World Congress on Transport Research*.

5.) Bartin, B., Ozbay, K., and Iyigun, C. A Clustered Based Methodology for Determining the Optimal Roadway Configuration of Detectors for Travel Time Estimation. Submitted to the 86th Transportation Research Board Annual Meeting. 2006.

6.) Beasley, J.E. and Christofides, N. (1989), An algorithm for the resource constrained shortest path problem, Networks, Vol. 9, No. 4, pp. 379-394.

7.) Bellman, R.E. (1958), On a Routing Problem. *Quart. Appl. Math. 16*, 87-90.

8.) Bianco, L., Confessore, G., and Reverberi, P. (2001), A network based model for traffic sensor location with implications on O/D matrix estimates, Transportation Science, Vol. 35, No. 1, pp. 50-60.

9.) Daganzo, C.F. (1994), The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory, Transportation Research Part B, Volume 28B, No. 4, pp. 269-287.

10.) Daganzo, C.F. (1995), The Cell Transmission Model, Part II: Network Traffic, Transportation Research Part B, Volume 29B, No. 2, pp. 79-93.

11.) Das, I. and Dennis, J.E. (1997), A Closer Look at Drawbacks of Minimizing Weighted Sums of Objectives for Pareto Set Generation in Multicriteria Optimization Problems, Structural Optimization, Vol. 14, no. 1, pp. 63-69.

12.) Das, I. and Dennis, J.E. (1998), Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems, SIAM Journal on Optimization, Vol. 8, No. 3, pp. 631-657.

13.) Denardo, E.V. and Fox, B.L. (1979), Shortest-route methods, 1. Reaching, pruning, and buckets, Operations Research 27, 161-186.

14.) Dial, R.B. (1969), Algorithm 360: Shortest path forest with topological ordering, Comm. ACM 12, 632-633.

15.) Dijkstra, E.W. (1959), A Note on Two Problems in Connection with Graphs, Numer. Math. 1, 269-271.

16.) Fei, X., Mahmassani, H.S., and Eisenman, S.M. (2007), Sensor Coverage and Location for Real-time Traffic Prediction in Large-Scale Networks, Transportation Research Record, Vol. 2039, pp. 1-15.

17.) Fei, X. and Mahmassani, H.S. (2007), A Two-Stage Stochastic Model for the Sensor Location Problem in a Large-Scale Network, Conference Paper, 87th Annual Transportation Research Board Meeting.

18.) Ford Jr., L.R. and Fulkerson, D.R. (1962), *Flows in Networks*, Princeton University Press, Princeton, NJ.
19.) Fujito, I., Margiotta, R., Huang, W., and Perez, W.A. (2006), The Effect of Sensor Spacing on Performance Measures. 85th Annual Transportation Research Board Meeting CD-ROM.
20.) Gendreau, M., Laporte, G., and Parent, I. (2000), Heuristics for the location of inspection stations on a network, Naval Research Logistics, Vol. 47, Issue 4, pp. 287-303.
21.) Handler, G.Y. and Zang, I. (1980), A dual algorithm for the constrained shortest path problem, Networks, Vol. 10, pp. 293-310.
22.) Jensen, P.A. and Berry, R.C. (1971), A constrained shortest path algorithm, Presented at the 39th National ORSA Meeting, Dallas, TX, USA.
23.) Joksch, H.C. (1966), The shortest route problem with constraints, Journal of Mathematical and Analytical Applications, Vol. 14, pp. 191-197.
24.) Kasprazak, E.M. and Lewis, K.E. (1999), A Method to Determine Optimal Relative weights for Pareto Solution Sets, Proceedings of the Third World Congress of Structural and Multidisciplinary Optimization (WCSMO-3) (Bloebaum, C.L., Lewis, K.E., et al., eds.), Buffalo, NY, University of Buffalo, Vol. 2, pp. 408-410.
25.) Kwon, J., Petty, K., and Varaiya, P. (2006), Probe Vehicle Runs or Loop Detectors? Effect of Detector Spacing and Sample Size on the Accuracy of Freeway Congestion Monitoring, Submitted for Presentation and Publication at the Transportation Research Board – 86th Annual Meeting.
26.) Li, Y., Fadel, G.M., and Wiecek, M.M. (1998), Approximating Pareto Curves Using the Hyper-Ellipse, 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, AIAA, Vol. 3, pp. 1990-2002.
27.) Liu, H. and Danczyk, A. (2007), Optimal Detector Placement for Freeway Bottleneck Identification, Conference Paper, 87th Annual Transportation Research Board Meeting.
28.) Moore, E.F. (1959), The Shortest Path through a Maze, *Proceedings of the Int. Symp. On the Theory of Switching*, Harvard University Press, 285-292.
29.) Ozbay, K., Bartin, B., and Chien, S. (2004), South Jersey Real-Time Motorist Information Systems: Technology and Practice. Transportation Research Record, 1886, pp. 68-75.
30.) Pape, U. (1974), Implementation and efficiency of Moore algorithms for the shortest root problem, Math. Prog. 7, 212-222.
31.) Rardin, R.L. (1998), Optimization in Operations Research, Prentice Hall, Inc., 440.
32.) Sherali, H.D., Desai, J., Rakha, H., and El-Shawarby, I. (2002), A Discrete Optimization Approach for Locating Automatic Vehicle Identification Readers for the Provision of Roadway Travel Times, Transportation Research Part B, Volume 40, Issue 10, pp. 857-871.
33.) Traffic Detector Handbook: Third Edition. United States Department of Transportation. Federal Highway Administration. Publication No. FHWA-HRT-06-108. October 2006.

34.) Watters, L. (1967), Reduction of Integer Polynomial Programming Problems to Zero-One Linear Programming Problems. Operations Research, Vol. 15, No. 6, pp. 1171-1174.
35.) Wilson, B., Cappelleri, D., Simpson, T.W., and Frecker, M. (2000), Efficient Pareto Frontier Exploration Using Surrogate Approximations, Proceedings of the 8th AAIA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Long Beach, CA.
36.) Witzgall, C. and Goldman, A.J. (1965), Most profitable routing before maintenance, Presented at the 27th National ORSA Meeting, Boston, MA, USA.
37.) Yang, H., Yang, C., and Gan, L. (2006), Models and algorithms for the screen line-based traffic-counting location problems, Computers and Operations Research, Vol. 33, Issue 3, pp. 836-858.
38.) Yang, H. and Zhou, J. (1998), Optimal traffic counting locations for origin-destination matrix estimation, Transportation Research Part B: Methodological, Vol. 32, Issue 2, pp. 109-126.
39.) Zhang, X. and Yang, H. (2004), The optimal cordon-based network congestion pricing problem, Transportation Research Part B, Vol. 38, Issue 6, pp. 517-537.

Flow model based density estimation method for highways using optimal sensor configurations

Olli-Pekka Tossavainen, Ryan Herring and Alexandre Bayen

February 24, 2009

1 Introduction

Density field along a highway is an important indicator of traffic state. It is used, for example, as input for ramp–metering algorithms that control the flow of vehicles that are allowed to enter the highway. In theory we would like to have as many loop detectors in the pavement as possible which would produce traffic measurements for density estimation. However, this is not possible in practice (costs money, physical limitations).

In this specific task of the project we aim to evaluate the performance of optimal sensor locations for density estimation on highways in the presence of a nonlinear Cell Transmission Model (CTM). The Cell transmission model provides a time evolution of a density field on a stretch of highway. However, the CTM model can benefit from existing loop detectors and also in the future from GPS equipped mobile phones reporting their speed, when these are combined in an estimation method.

The reason behind the use of CTM for providing the time evolution of the density is based on the fact that the sensor coverage is not often very dense along the highway or the sensors cannot always be placed in best locations. Also some of the sensors may be broken. In these cases it is crucial to get the information between the sensors from a model that has been validated to produce reliable estimates of density in the areas where there are no direct sensor measurements. Simple interpolation methods are not sufficient to predict the density and they rarely have any acceptable time dependent behavior.

The estimation method that we use in this study is based on a special Kalman filtering technique called Ensemble Kalman Filtering (EnKF) [4]. It is a special technique that does not require any mode knowledge of the state of the highway (free flow, congested, shock wave present) and thus can use a fully nonlinear CTM model, in contrast to studies conducted for example in [6, 8].

The main contribution of this task was to evaluate the performance of the CTM based estimation method in the case of optimized sensor configuration versus a naive density estimation method where sensor readings are used directly to define the density field on a highway. The results show that in general it is better to use a flow model as a basis for density estimation instead of direct interpolation between sensor readings. This is also intuitive since the dynamics of the highway are very nonlinear and thus interpolation methods can lead to misinterpretation of the sensor readings.

The performance of the different sensor configurations was also assessed by performing estimation with randomly generated sensor locations and using CTM. The results from these experiments suggest that on the average CTM based estimation produces comparable or even better results than a naive estimation method with optimized sensor configuration.

The optimum for the sensor locations can be found for example using travel time as a cost function as done in [2]. This is the criteria used in this case study. Criteria can also be minimization of the error in density estimation (which is a possible long term goal).

The rest of this section is organized as follows: In Section 2 we introduce the basic concept of flow model based density estimation. In Section 3, demonstrate the performance of the estimation with optimized sensor locations using a varying number of sensors. In Section 4, we compare the performance of randomly spaced sensors versus optimized sensor locations These results serve a basis for concluding how sensor configuration obtained from different criteria can be used in other applications.

2 Density estimation framework

The problem is formulated using state–space formalism. The density of vehicles on a given stretch of highway is assumed to obey cell transmission model (CTM) such that

$$\rho_{k+1} = F_k(\rho_k) + w_k. \tag{1}$$

Here ρ_k is the predicted system state at time step k and $F_k(\rho_k)$ is one time step in the (non-linear) CTM. Formula (1) is usually called the state evolution equation.

For the observations, we use an additive noise model

$$y_k = H_k\rho_k + \epsilon_k. \tag{2}$$

The observation vector is y_k and the observation model that relates the state variables to the measurements is H^n. The noise process ϵ_k is the measurement noise with covariance R_k. Without loss of generality, noise processes are assumed to have zero mean.

2.1 Density CTM

The cell transmission model employed in this study is based on the implementation presented in [7]. More specifically, the density on a highway evolves according to a finite difference scheme

$$\rho_{k+1}^{(i)} = \rho_k^{(i)} - r\left(q\left(\rho_k^{(i+\frac{1}{2})}\right) - q\left(\rho_k^{(i-\frac{1}{2})}\right)\right) \tag{3}$$

where subscript k and superscript i refer to time and space discretization, respectively. Furthermore, r is the ratio between desired space and time discretization and q is the so-called Godunov flux. The flux function encodes the information from critical density, jam density etc. into a usable form for the numerical method. These variables indicate for example the maximum capacity of the highway and the critical density ρ^{crit} (typically 33-45 veh/mile/lane).

For the concave flux function q the numerical Godunov flux q_{G} can be written as

$$q_{\text{G}}(\rho^{(1)}, \rho^{(2)}) = \begin{cases} q(\rho^{(2)}) & \text{if } \rho^{\text{crit}} < \rho^{(2)} < \rho^{(1)} \\ q(\rho^{\text{crit}}) & \text{if } \rho^{(2)} < \rho^{\text{crit}} < \rho^{(1)} \\ q(\rho^{(1)}) & \text{if } \rho^{(2)} < \rho^{(1)} < \rho^{\text{crit}} \\ \min(q(\rho^{(1)}), q(\rho^{(2)})) & \text{if } \rho^{(1)} \leq \rho^{(2)} \end{cases}$$

Due to the severe non-linearity in the flux function, a special technique to incorporate measurements from loop detector into this flow model is required. For complete details of the scheme and the employed flux function, reader is referred to [7].

The underlying model has been shown to predict densities on a highway accurately without using any estimation techniques on top of it. However, these cases typically don't consider effects

of on-ramps and off-ramps. On long stretches of highways under consideration these cause severe errors in the density field produced by the CTM model. In this study the loop detectors are used to compensate this error source by incorporating their reported measurements into the predicted density of CTM. Also the proposed framework tracks the error in the estimation procedure which cannot be achieved by just running the CTM alone.

2.2 EnKF algorithm

In the filtering problem, the aim is to compute conditional expectations

$$\rho_{k|k} = \mathbb{E}(\rho_k | y_k, \ldots, y_1).$$

In the case of linear observation and evolution equations and Gaussian noise processes, the recursive Kalman filter algorithm can be used for determining the estimates of conditional expectation $\theta_{k|k}$ and covariance $\Gamma_{k|k}$. If evolution and/or observation equations are nonlinear, the Extended Kalman Filter (EKF) can be applied [1]. The use of the EKF also requires that the operator $F_k(\cdot)$ be differentiable. However, to avoid the difficult linearization of the model F and to preserve the higher order statistics, which may be lost in the linearization, we employ the Ensemble Kalman filter.

For the state space model (1)–(2), the EnKF algorithm can be summarized as in [5, 4]:

1. Initialization: An ensemble of N states $\xi_0^{(i)}$ are generated to represent the uncertainty in ρ_0.
2. Time update:

$$\xi_{k|k-1}^{(i)} = F(\xi_{k-1|k-1}^{(i)}) + w_{k-1}^{(i)} \tag{4}$$

$$\theta_{k|k-1} = \frac{1}{N}\sum_{i=1}^{N} \xi_{k|k-1}^{(i)} \tag{5}$$

$$E_{k|k-1} = [\xi_{k|k-1}^{(1)} - \theta_{k|k-1}, \ldots, \xi_{k|k-1}^{(N)} - \theta_{k|k-1}] \tag{6}$$

3. Measurement update:

$$\Gamma_{k|k-1} = \frac{1}{N-1} E_{k|k-1} E_{k|k-1}^{\mathrm{T}} \tag{7}$$

$$K_k = \Gamma_{k|k-1} H_k^{\mathrm{T}} [H_k \Gamma_{k|k-1} H_k^{\mathrm{T}} + R_k]^{-1} \tag{8}$$

$$\xi_{k|k}^{(i)} = \xi_{k|k-1}^{(i)} + K_k [y_k - H_k \xi_{k|k-1}^{(i)} + \epsilon_k^{(i)}] \tag{9}$$

where the ensemble of state vectors are generated with the realizations $w_k^{(i)}$ and $\epsilon_k^{(i)}$ of the noise processes w_k and ϵ_k, respectively. In the previous equations, an important step is that at measurement times, each measurement is represented by an ensemble. This ensemble has the actual measurement y_k as mean and the variance of the ensemble is used to represent measurement errors. This is done by adding perturbations $\epsilon_k^{(i)}$ to measurements drawn from a distribution with zero mean and covariance equal to measurement error covariance matrix R_k. This ensures that the updated ensemble has a variance that is not too low [3].

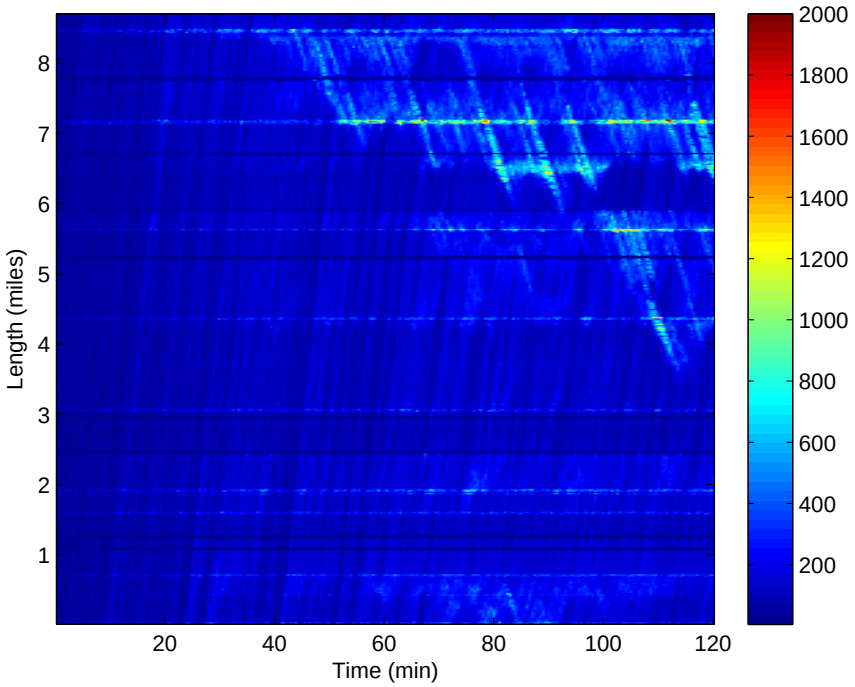


Figure 1: Example of a density field obtained from Paramics micro simulation data. Color scale represents the density of vehicles per mile on the highway.

3 Optimal sensor location based estimation

The test site consists of 45900 ft long segment on Interstate 880 North bound between post miles 17.7 and 26.4 (between Mowry Ave and Tennyson Rd). Six different Paramics (micro simulator) simulations are performed to get the ground truth density fields and to simulate the sensor density data. The selected time scale represents a typical two-hour-period during the evening commute. An example of the Paramics simulation produced density field is given in Figure 1.

The results are compared between two different density estimation methods.

1. Naive estimation method: Each sensor is assigned with an area of influence. The area of influence can be for example defined between the midpoints of neighboring sensor locations. Throughout this area the density is assumed to be constant (as read by the sensor) until the next sensors area of influence is reached. This is a commonly used intuitive method in traffic modeling. In the worst cases this area of influence can be even 0.5 miles.

2. CTM based data assimilation method that employs ensemble Kalman filter algorithm. In this novel algorithm (never employed for highway density estimation to our best knowledge) we use the CTM model to evolve the dynamics of the traffic and use the loop detector data to provide additional information of the traffic. The highway is discretized into 900 feet cells and this forms the resolution of the model. Sensors are mapped to the corresponding cell of the discretization and it is assumed that sensor reading is a measurement of the density of this particular cell.

The results are given as a relative error. More, specifically we show how well each method can

detect congestion on the highway. We check if the cell is congested according to formula

$$\rho_{\text{binary}}(\rho) = \begin{cases} 0 & \text{if } \rho \leq \rho^{\text{crit}} \\ 1 & \text{if } \rho > \rho^{\text{crit}} \end{cases}$$

By doing this way the performance of the method shows its capability of estimating the lengths of the queues that are formed on the highway, instead of just penalizing the errors in the individual density values. This corresponds to typical binning of speeds into specific categories when presenting the speed contours on highway.

We present the results that compare the performance of the naive estimation method and CTM based estimation in the case of optimal sensor placement. The optimal sensor locations are computed using the algorithm [2] (also developed in this project).

In Figure 2 relative errors for the naive interpolation based method and EnKF method are presented. By looking at all seven cases, it can be concluded that in general flow model based estimation method always outperforms the naive interpolation method. This is due, for example, to the fact that the presented estimation frame work includes the actual physics from the highway and is able to handle measurement errors due to its formulation. The naive interpolation does not treat the extent of the backward propagating congestion in physical manner and thus causes misinterpretation of the density along the highway. The naive estimation method is also very sensitive to the place of the sensor since it assigns the sensor reading as is to a relatively long segment of the highway in contrast to the CTM based method.

4 Random sensor location based estimation

In this section we compare the results of CTM model based estimation with random sensor placement to a naive approach with optimized sensor locations to give an idea of the power of the flow model based estimation. The sensors get placed almost evenly for the CTM model along the highway and their locations are then perturbed around their initial placement still maintaining the order of the sensors.

Results are presented in Figure 3. The results can be interpreted such that even randomly placing the sensors on the highway and using the CTM based estimation, better results are generally achieved than using optimized sensor locations and naive method. This is based on the fact that the performance of the naive density estimation method is heavily dependent on the success of the sensor placement. Even when the sensors are not optimally placed the CTM based method captures the main patterns of the traffic better than naive "area of influence" based interpretation of the sensor measurements.

The results of CTM estimation with optimal configuration that are presented in Figure 3 generally fall within the standard deviation of the error in Figure 2.

These results do not indicate that it would be meaningless to optimize sensor locations, instead by carefully placing the sensors for travel time estimation (interest of the driving public), the same sensor configuration can be used to produce good results in density estimation as long as the estimation method is chosen wisely.

5 Notes

The results presented above demonstrate the power of flow model based estimation. The optimal sensor configuration is provided by optimization method developed also in this project and is based

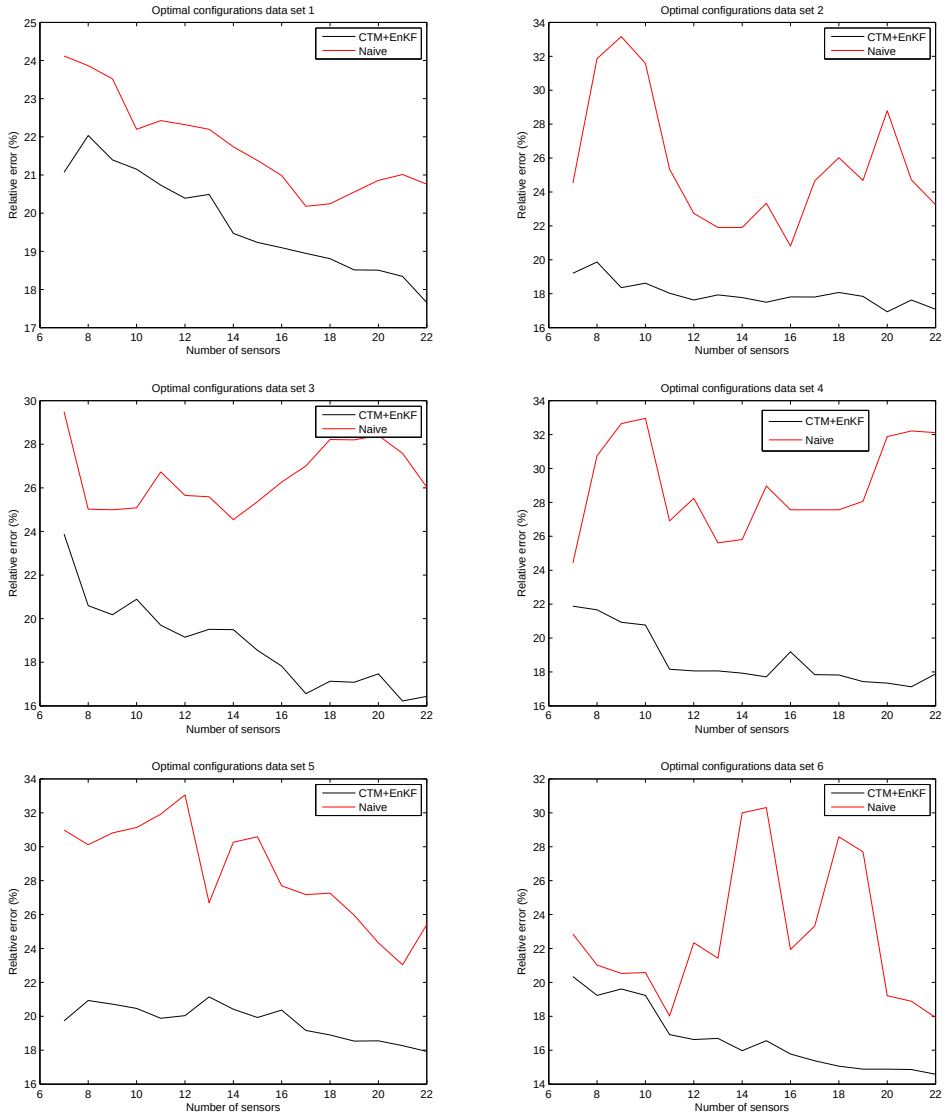


Figure 2: Relative errors in the estimated density field for six different cases for optimal sensor locations. Results are shown for CTM based density estimation and naive interpolation method.

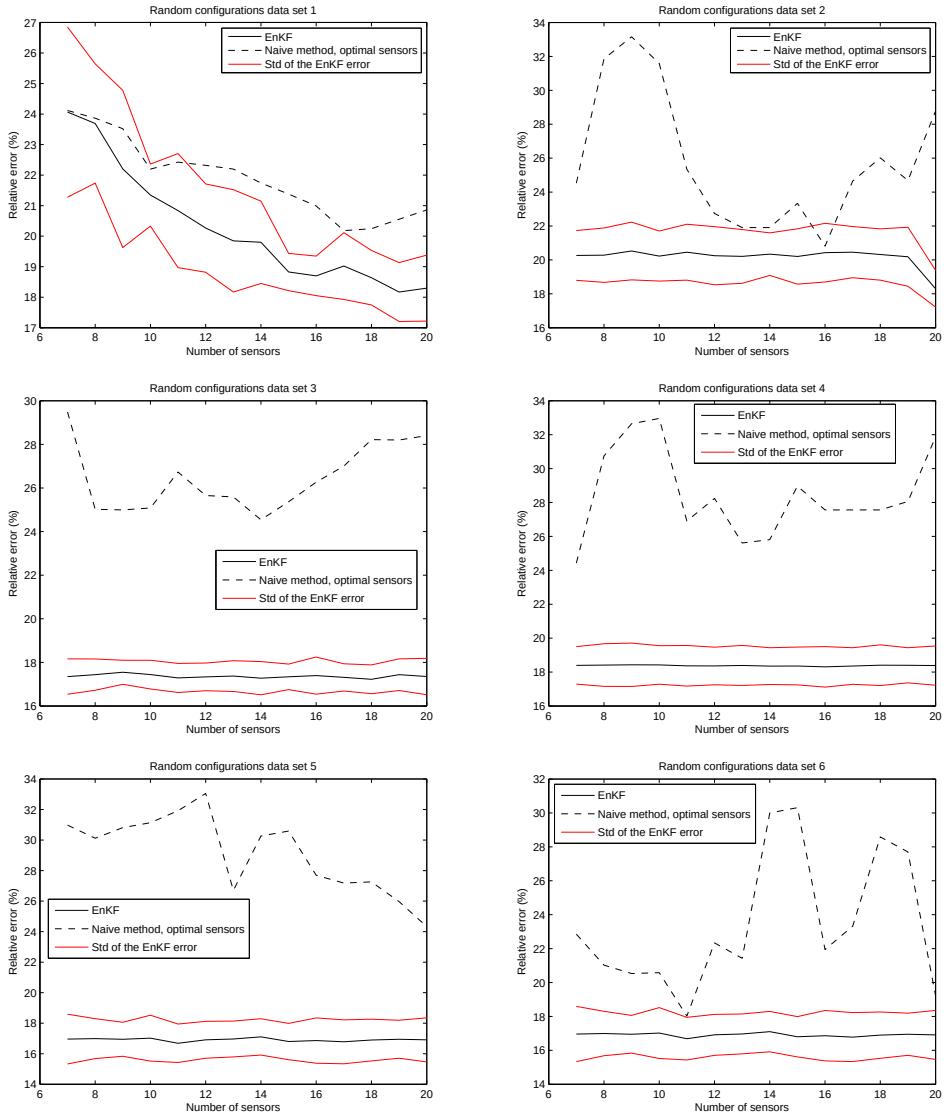


Figure 3: Relative errors in the estimated density field for six different cases for random sensor locations. Results are shown for CTM based density estimation and naive interpolation method (using optimal configuration).

on optimization over travel time computation [2]. The optimal sensor placement for purely density estimation is likely to be different than for travel time estimation purposes due to different underlying dynamics on the highway. However, the optimization using travel time is mathematically simpler than using the quality of the density estimation as a criteria. Thus, the employed optimization scheme for sensors is very tractable.

It has already been shown in this project that by optimizing the locations for travel time, a significant improvement in the travel time estimation is achieved. The same optimized sensor locations provide also feasible locations for density estimation, although they may not be the most optimal because of the complexity of the density evolution dynamics. However, the travel time study can be easily deployed to almost any site without setting up a relatively heavy machinery to optimize using density estimation as a criteria. Using these travel time optimized locations one can still have practical value of using these same sensor locations when performing density estimation. The risks of deploying a poor random sensor configuration also for the density estimation are big when only limited number of sensors to be installed are available. Thus, it is always preferable to use a optimization method when considering placement of sensors.

There is a currently ongoing research that aims to deploy the optimal placement (with maximum spacing) of virtual loop detectors in Mobile Millennium project. The deployment is always a compromise between cost (bandwidth, infrastructure) and detectability of the traffic. Also, due to the nature of traffic dynamics one would in theory want to move the sensors based on AM/PM rush hours and during accidents.

References

[1] B.D.O. Anderson and J.B. Moore. *Optimal filtering*. Prentice-Hall, inc, 1979.

[2] X. Ban, R. Herring, J. D. Margulici, and A. Bayen. Optimal sensor placement for freeway travel time estimation. *Submitted to the 18th International Symposium on Transportation and Traffic Theory*, 2008.

[3] G. Burgers, P. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724, 1998.

[4] G. Evensen. *Data Assimilation – The Ensemble Kalman Filter*. Springer, 2007.

[5] A.W. Heemink, M. Verlaan, and A.J. Segers. Variance reduced ensemble Kalman filtering. *Monthly Weather Review*, 129:1718–1728, 2001.

[6] L. Munoz, X. Sun, R. Horowitz, and L. Alvarez. Traffic density estimation with the cell transmission model. In *Proceedings of the 2003 American Control Conference*, Denver, Colorado, June 4-6 2003.

[7] I. Strub and A.M. Bayen. Weak formulation of boundary conditions for scalar conservation laws: an application to highway modeling. *International Journal on Robust and Nonlinear Control*, 16:733–748, 2006.

[8] X. Sun, L. Munoz, and R. Horowitz. Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawaii, December 9-12 2003.

SECTION 6 – OPTIMAL DETECTOR PLACEMENT TOOLKIT

A. DANCZYK, H. LIU. “SENSOR ALLOCATION PROGRAM FOR FREEWAY CORRIDOR”. DEPARTMENT OF CIVIL ENGINEERING, UNIVERSITY OF MINNESOTA, 2009.

P. MCGOWEN. “RURAL ISSUES WITH OPTIMAL SENSOR PLACEMENT FOR TRANSPORTATION APPLICATIONS”. WESTERN TRANSPORTATION INSTITUTE, 2008.

Sensor Allocation Program for Freeway Corridor

INSTRUCTION MANUAL

Adam Danczyk

Henry X. Liu

Department of Civil Engineering

University of Minnesota

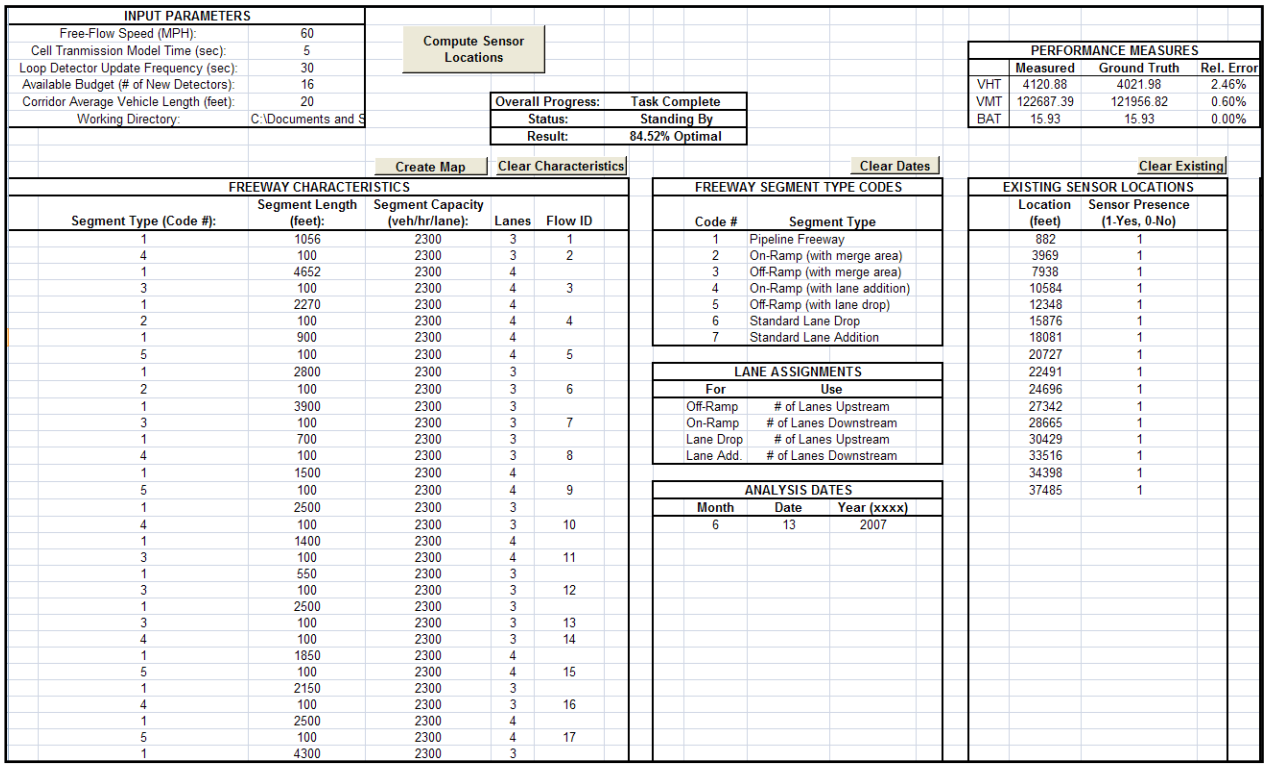
March 2009

Table of Contents

Introduction

The 'Sensor Allocation Program' is an intuitive tool designed to aid engineers in determining sensor allocations on a freeway corridor. Using the allocation model developed by Liu and Danczyk (2009), the program optimizes sensor placement to maximize bottleneck detection accuracy, as well as compute a few popular performance measure accuracies that would result from such a deployment. Additionally, it offers a repair priority schedule for engineers to better determine which sensors to repair given multiple failures and a limited budget. This tool is applicable to corridors that lack instrumentation as well as corridors with existing sensor infrastructure.

The purpose of this document is to guide the user through using this tool. It discusses the inputs necessary and the formatting needed to allow the tool to run correctly. Furthermore, it offers interpretation of the outputs following the completion of the model's run. Figure 1 shows the main page of the program.



INPUT PARAMETERS

Free-Flow Speed (MPH):	60
Cell Tranmission Model Time (sec):	5
Loop Detector Update Frequency (sec):	30
Available Budget (# of New Detectors):	16
Corridor Average Vehicle Length (feet):	20
Working Directory:	C:\Documents and S

Compute Sensor Locations

Overall Progress:	Task Complete
Status:	Standing By
Result:	84.52% Optimal

PERFORMANCE MEASURES

	Measured	Ground Truth	Rel. Error
VHT	4120.88	4021.98	2.46%
VMT	122687.39	121956.82	0.60%
BAT	15.93	15.93	0.00%

Create Map | Clear Characteristics | Clear Dates | Clear Existing

FREEWAY CHARACTERISTICS

Segment Type (Code #):	Segment Length (feet):	Segment Capacity (veh/hr/lane):	Lanes	Flow ID
1	1056	2300	3	1
4	100	2300	3	2
1	4652	2300	4	
3	100	2300	4	3
1	2270	2300	4	
2	100	2300	4	4
1	900	2300	4	
5	100	2300	4	5
1	2800	2300	3	
2	100	2300	3	6
1	3900	2300	3	
3	100	2300	3	7
1	700	2300	3	
4	100	2300	3	8
1	1500	2300	4	
5	100	2300	4	9
1	2500	2300	3	
4	100	2300	3	10
1	1400	2300	4	
3	100	2300	4	11
1	550	2300	3	
3	100	2300	3	12
1	2500	2300	3	
3	100	2300	3	13
4	100	2300	3	14
1	1850	2300	4	
5	100	2300	4	15
1	2150	2300	3	
4	100	2300	3	16
1	2500	2300	4	
5	100	2300	4	17
1	4300	2300	3	

FREEWAY SEGMENT TYPE CODES

Code #	Segment Type
1	Pipeline Freeway
2	On-Ramp (with merge area)
3	Off-Ramp (with merge area)
4	On-Ramp (with lane addition)
5	Off-Ramp (with lane drop)
6	Standard Lane Drop
7	Standard Lane Addition

LANE ASSIGNMENTS

For	Use
Off-Ramp	# of Lanes Upstream
On-Ramp	# of Lanes Downstream
Lane Drop	# of Lanes Upstream
Lane Add.	# of Lanes Downstream

ANALYSIS DATES

Month	Date	Year (xxxx)
6	13	2007

EXISTING SENSOR LOCATIONS

Location (feet)	Sensor Presence (1-Yes, 0-No)
882	1
3969	1
7938	1
10584	1
12348	1
15876	1
18081	1
20727	1
22491	1
24696	1
27342	1
28665	1
30429	1
33516	1
34398	1
37485	1

Figure 1: Main Page

Requirements

This tool was designed for and intended to be used in Microsoft Excel 2003. It was written in the .xls format and can be successfully opened in later versions of Microsoft Excel that allow for this file type. It is strongly advised that a backup copy be made of this tool should the document become corrupted.

This tool requires loop detector data of the corridor in question to simulate traffic. This data needs to be for the upstream access point (on freeway) and for all on-ramps and off-ramps. No data is required other than that. Data needs to be stored in a folder labeled as **Loop Detector Data**. Data files need to be given in the format of 'M-D-YYYY.csv' in this folder, where M is month (1-12), D is date (1-31), and YYYY is year (2007, 2008, 2009, etc.). The location of this 'Loop Detector Data' folder is considered the Working Directory and it is recommended to place the program file in this directory. Failure to correctly do this will result in the program not functioning properly.

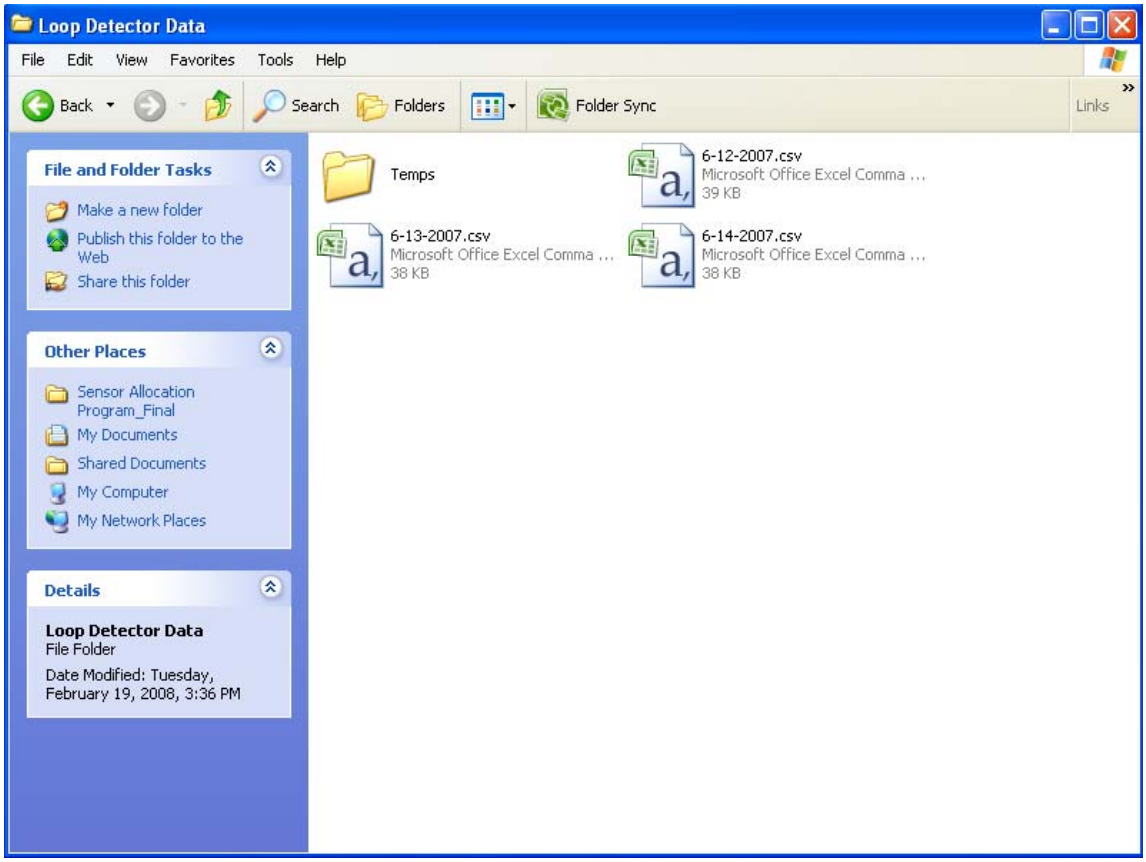


Figure 2: Loop Detector Data Folder with three days of data.

General Layout

This tool is subdivided over three Excel tabs, which can generally be found at the bottom of the screen. The three tabs include:

- **Input Parameters**: The page where data regarding freeway characteristics is entered and output data, in numerical values, can be calculated. This page is where most of the work for this tool is conducted.
- **Map**: The page where a map representation of the calculated result can be generated for visual understanding. Use of this page requires data to be previously calculated on the 'Input Parameters' page.
- **Stored Data**: A miscellaneous page where extra data can be stored, such as freeway characteristics intended to be used at a later time. The data placed on this page does not impact the operation of the tool, nor can the data on this page alone be used to operate the tool without being moved to the 'Input Parameters' page.

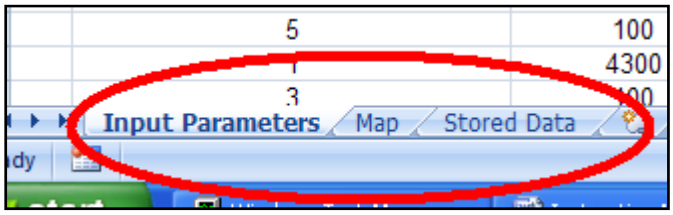


Figure 3: The Three Tabs in Bottom Left Corner of Microsoft Excel

With most of the critical components to this project being found under the 'Input Parameters' page, the remainder of discussion on the general layout will focus on that page. Details on the 'Map' page can be found later in the document. Since 'Stored Data' is more intended as a storage site, it will not be discussed in any further detail.

On the 'Input Parameters' page, 10 boxes can be found around the page, as well as 8 buttons. Each box and button serves a distinct purpose. The 10 boxes include:

- **Input Parameters**: Various general parameters are entered in this box, including Free-Flow Speed, Cell Transmission Model Step Time, Loop Detector Update Frequency, Available Sensor Budget, Average Vehicle Length, and Working Directory. These apply to the entire corridor.
- **Freeway Characteristics**: The characteristics of the freeway, divided into characteristic-based segments, are defined in this box. Each segment includes a given segment length, capacity, number of lanes, and presence of a flow (if applicable). Characteristics are determined on a segment-by-segment basis.

- **Overall Progress**: This box, situated above ‘Freeway Characteristics’ and ‘Freeway Segment Type Codes’, shows progress of the model when running and some end result data.
- **Freeway Segment Type Codes**: This box gives codes for the ‘Freeway Characteristics’ box to aid in determining what type of segment is being used. Basic segments include pipelines, on-ramps (with or without lane additions), off-ramps (with or without lane additions), lane drops, and lane additions.
- **Lane Assignments**: This box aids in determining how many lanes to use at sites where on-ramps or off-ramps are present.
- **Analysis Dates**: A user-defined date of analysis.
- **Performance Measures**: This box informs the user, after the model completes its analysis, of several performance measures that are expected to result from the given sensor configuration, as well as their accuracies. These measures include Vehicle-Hours Traveled (VHT), Vehicle-Miles Traveled (VMT), and Bottleneck Activity Time (BAT).
- **Existing Sensor Locations**: Used to define where sensors presently exist on the corridor.
- **New Configuration**: Following the completion of the model’s operation, the recommendations of sensor allocations are placed here for user reference.
- **Repair Importance**: Following the completion of the model’s operation, the recommendations of repair priority for these sensors is given, should future failure occur and the user’s budget is limited.

Among the 8 buttons present on this page, the purposes of each are:

- **Compute Sensor Locations**: Executes the computation process for optimizing sensor locations.
- **Create Map**: Generates a map under the ‘Map’ tab based on the freeway characteristics.
- **Clear Characteristics**: Clears the characteristics defined in the ‘Freeway Characteristics’ box. Will also clear recommendations for a new configuration.
- **Clear Dates**: Clears the dates listed under ‘Analysis Dates’. Will also clear recommendations for a new configuration.
- **Clear Existing**: Clears the detectors listed under ‘Existing Sensor Locations’. Will also clear recommendations for a new configuration.
- **View Allocation**: Generates a map under the ‘Map’ tab based on the freeway characteristics and new sensor configuration.
- **Clear New Sensors**: Clears the detectors listed under ‘New Configuration’.
- **Clear Repair**: Clears the detectors and priorities under ‘Repair Importance’.

These buttons and boxes will play an important role when preparing to allocate sensors in the following sections.

Inputting Model Parameters

In this section, it will be discussed how to prepare characteristics of the corridor in question for use with the model. These parameters are needed to operate the traffic simulator, which in this program is the Cell Transmission Model proposed by Daganzo (1994) and Daganzo (1995).

Input Parameters Box

INPUT PARAMETERS	
Free-Flow Speed (MPH):	60
Cell Tranmission Model Time (sec):	5
Loop Detector Update Frequency (sec):	30
Available Budget (# of New Detectors):	16
Corridor Average Vehicle Length (feet):	20
Working Directory:	C:\Documents and S

Figure 4: Input Parameters Box

To begin, corridor-wide parameters need to be entered. Under the 'Input Parameters' box, the following data needs to be entered:

1. **Free-Flow Speed**: Approximate what the corridor-wide free-flow speed is. Under most situations, this will vary with different segments, but limitations with the cell transmission model require a constant value to be inputted. In this program, English units will be used. Generally, the free-flow speed is taken as 5 miles per hour (MPH) higher than the speed limit. A recommended value would be 60 MPH (as to correspond with general urban freeway speeds of 55 MPH).
2. **Cell Transmission Model Time**: Determine what time interval will be used for the cell transmission model. This time interval needs to have a whole-number multiple with the Loop Detector Update Frequency (discussed later). A recommended value would be 5 seconds. Any smaller will render very small cells and is not advised. Any larger will result in broader accuracies.
3. **Loop Detector Update Frequency**: Determine how often loop detector data used for simulation updates, in seconds. In general, this value is about 30 seconds, but can be changed depending on how the data was originally downloaded. This value needs to be a whole-number multiple of the Cell Transmission Model Time. That is, if the Loop Detector Update Frequency is 30 seconds, the Cell Transmission Model Time should be, for example, 5 seconds, 10 seconds, or 15 seconds, and **not** 20 seconds or 25 seconds. This is very important. Failing to do so will cause an error or an incorrect approximation.
4. **Available Budget**: This parameter is the number of detectors available for deployment. In this program, it is assumed that detector deployment at any given site comes with the same uniform cost and that all detectors made available are deployed. If more sensors are

budgeted than there is space along the corridor, a maximum number will be reassessed and deployed.

5. **Corridor Average Vehicle Length**: Determine the average vehicle length for vehicles traveling along the corridor, in feet. A recommended value would be 20 feet.
6. **Working Directory**: Define the directory in which the loop detector data (separated by date) is located. For example, if the data file is located on the desktop, the directory usually would be "C:\Desktop\". Make sure the last character in this name is a "\". Additionally, make sure the 'Loop Detector Data' folder holding the loop detector data is in the same directory.

These parameters can only be deleted one at a time by the user.

Freeway Characteristics Box

Create Map | Clear Characteristics

FREEWAY CHARACTERISTICS

Segment Type (Code #):	Segment Length (feet):	Segment Capacity (veh/hr/lane):	Lanes	Flow ID
1	1056	2300	3	1
4	100	2300	3	2
1	4652	2300	4	
3	100	2300	4	3
1	2270	2300	4	
2	100	2300	4	4
1	900	2300	4	
5	100	2300	4	5
1	2800	2300	3	
2	100	2300	3	6
1	3900	2300	3	
3	100	2300	3	7
1	700	2300	3	
4	100	2300	3	8
1	1500	2300	4	
5	100	2300	4	9

Figure 5: Freeway Characteristics Box

The next step is to define characteristics of each segment across the corridor. In the 'Freeway Characteristics' box, these definitions can be made for each segment. The number of segments depends on the size of the Excel document and can often range up to 65,000, although such a number is not recommended because of time and memory constraints. When first starting, use the first row below the column headers. Steps for defining the characteristics are listed below:

1. **Define the Segment Type:** Under the column labeled 'Segment Type (Code #)', select the code for the segment type. This code can be found in the box to the right labeled 'Freeway Segment Type Codes'. These segments codes are also given here (**1** – Pipeline Freeway, **2** – On-Ramp (with merge area), **3** – Off-Ramp (with merge area), **4** – On-Ramp (with lane addition), **5** – Off-Ramp (with lane drop), **6** – Standard Lane Drop, **7** – Standard Lane Addition).

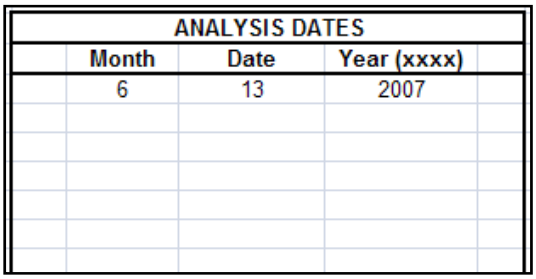
FREEWAY SEGMENT TYPE CODES	
Code #	Segment Type
1	Pipeline Freeway
2	On-Ramp (with merge area)
3	Off-Ramp (with merge area)
4	On-Ramp (with lane addition)
5	Off-Ramp (with lane drop)
6	Standard Lane Drop
7	Standard Lane Addition

Figure 6: Freeway Segment Type Codes Box

2. **Define the Segment Length:** Under the column labeled "Segment Length (feet)", define the segment length, in feet.
3. **Define the Segment Capacity**: Under the column labeled "Segment Capacity (veh/hr/lane)", define the segment capacity, in vehicles per hour per lane. A typical value for freeways is 2300 vehicles per hour per lane.
4. **Define the Number of Lanes**: Under the column labeled "Lanes", define the number of lanes. At areas of interchanges or lane number adjustments, refer to the 'Lane Assignments' box. Generally, for lane drops or off-ramps, the number of lanes equals the number of lanes upstream of the site. For lane additions or on-ramps, the number of lanes equals the number of lanes downstream of the site.
5. **Define Inbound/Outbound Traffic Flows**: Under the column labeled "Flow ID", define the identifying value for a given flow characteristic. For this program, flows only occur at the first segment (inbound flows from upstream), on-ramps (flow added), and off-ramps (flow removed). First segment flows are assessed through a data source on the freeway, such as an existing loop detector. On-ramp and off-ramp flows are assessed through data sources on the ramps themselves. It is very important to get these values correct, as otherwise the number of vehicles on freeway segments will be incorrect. This is discussed in more detail at a later section.
6. **Proceed to Next Segment, if applicable**: If another segment exists on the corridor of question, repeat steps 1-5 for that segment in the next row. If not, this task is finished.

If it is desired to clear this data, the 'Clear Characteristics' button can be used. Similarly, if a map of the freeway is desired, the 'Create Map' will do so.

Analysis Dates Box



ANALYSIS DATES		
Month	Date	Year (xxxx)
6	13	2007

Figure 7: Analysis Dates Box

The next step is to define the data intended to be used. Loop detector data must be stored in .csv format in a directory that can be defined in the 'Working Directory' row of the 'Input Parameters' box. The proper format for this file name is M-D-YYYY, such as 6-8-2007, 5-13-2007, or 11-28-2007. Similarly, in the 'Analysis Dates' box, such data needs to be entered in the same format. Starting in the first row of this box, steps for completing this are listed below:

1. **Define the Month**: Under the column labeled 'Month', enter the numerical month value, M, from the file. Using the three examples given above, such months would be 6, 5, or 11, respectively.
2. **Define the Date**: Under the column labeled 'Date', enter the numerical date value, D, from the file. Using the three examples given above, such dates would be 8, 13, or 28, respectively.
3. **Define the Year**: Under the column labeled 'Year (xxxx)', enter the numerical year value, YYYY, from the file. For all three examples given above, the year would be 2007.
4. **Proceed to Next Date, if applicable**: If another date is intended to be simulated, repeat steps 1-3 for that date in the next row. If not, this task is finished.

Note: All time values in the Analysis Date file need to be sequentially in equal intervals. All data in these files will be simulated.

If it is desired to clear this data, the 'Clear Dates' button can be used.

Existing Sensors Box

Clear Existing

EXISTING SENSOR LOCATIONS	
Location (feet)	Sensor Presence (1-Yes, 0-No)
882	1
3969	1
7938	1
10584	1
12348	1
15876	1
18081	1
20727	1
22491	1
24696	1
27342	1
28665	1
30429	1
33516	1
34398	1
37485	1

Figure 8: Existing Sensor Locations Box

If other loop detectors exist on the corridor with locations that cannot be changed and the user wishes to deploy additional loop detectors to compliment these existing ones, these existing ones can be defined. If no existing sensors are present, this step can be skipped. Starting with the first row beneath the column headers in the 'Existing Sensor Locations' box, steps for identifying existing detectors are listed below:

1. **Define Location of Existing Sensor**: Identify the stationing (in feet from beginning of analysis area) of the existing sensor. Place this value beneath the column labeled 'Location'. For example, if the existing sensor was 500 feet from the corridor observed, then the location would be 500. Identify locations sequentially.
2. **Define Presence of an Existing Sensor**: Under the column labeled 'Sensor Presence', place a value of 1 in rows with existing sensor, else place a 0 where no existing sensor is present. This is useful for changing the presence of existing sensors without having to rewrite the entire list.
3. **Proceed to the Next Existing Sensor, if applicable**: If another existing sensor is intended to be included, repeat steps 1-2 for that sensor in the next row. If not, this task is finished. Remember to make the next sensor be the next one in sequence.

If it is desired to clear this data, the 'Clear Existing' button can be used.

Running the Model

Once the parameters have been properly entered into the program, the program is ready to allow the model to optimize sensor locations. Getting the model to start running only requires the user to press the 'Compute Sensor Locations' button.

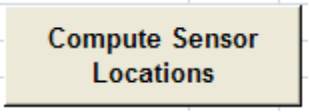


Figure 9: Compute Sensor Locations Button

Pressing this button will cause the 'Overall Progress' box to begin showing assessment. The program conducts several tasks as it runs from start to finish. These tasks are outlined below:

1. Assess Freeway Corridor Characteristics
2. Simulate Traffic Using Cell Transmission Model
3. Use a Genetic Algorithm to Optimize Sensor Locations
4. Compute Performance Measures
5. Determine the Priority of Each Sensor

In the 'Overall Progress' box, three rows reveal the progress of these tasks.

- **Overall Progress**: This shows the overall progress for the first four tasks, in terms of percentage completed.
- **Status**: This shows the current process being run. It will cycle through 'Estimating Traffic', 'Processing Algorithm', 'Calculating Performance Measures', and 'Prioritizing Repair' for Tasks 2 through 5, respectively. When the process is complete, it will say 'Standing By'.
- **Result**: This shows either that a result is being calculated or the end result. When Tasks 2 through 4 are running, it will be 'Calculating'. When Task 5 is running, it will be 'Prioritizing". When all of these tasks are completed, it will produce a percentage of optimality. This percentage represents the benefit captured, defined by the value of benefit in Liu and Danczyk (2009), relative to the benefit captured for a complete deployment. Thus, if it reveals that the result is 80 percent optimal, then the benefit captured could, say, be 800 while a deployment at all sites would have captured 1000.

Overall Progress:	67.53%
Status:	Processing Algorithm
Result:	Calculating...

Figure 10: Overall Progress Box when Program is Running

Troubleshooting

If an error is present that prevents the program from operating successfully, either due to a problem in data entry or source identification, the program will terminate and try to identify the cause. Generally, an error has occurred when the overall progress is identified as 'Task Terminated'. The 'Status' row in the box will aid in determining where the problem is located. The known problems are listed below:

- **Status: Missing Parameters** → Values are missing from the 'Input Parameters' box. Go back to the 'Input Parameters' box and enter the values into their correct box.
- **Status: Missing Road Info** → No road has been defined in the 'Freeway Characteristics' box. Check that segments have been entered into this box and that the first segment starts immediately below the column info (no rows should separate the two).
- **Status: Need Pipeline at Start** → The first segment defined in the 'Freeway Characteristics' box needs to be a pipeline segment. If the corridor does not start with a pipeline, this can be remedied by adding a 1-foot pipeline segment at the beginning of the corridor.
- **Status: Missing Date Info** → No date for data has been defined in the 'Analysis Dates' box. Please define the date or dates and ensure that the first date falls immediately below the column headers (no rows should separate the two).
- **Status: Need Pipeline at End** → The last segment defined in the 'Freeway Characteristics' box needs to be a pipeline segment. If the corridor does not end with a pipeline, this can be remedied by adding a 1-foot pipeline segment at the end of the corridor.
- **Status: Existing Location Error** → Given the defined Cell Transmission Model Time in the 'Input Parameters' box, the proximity of two existing sensors defined in the 'Existing Sensors' box result in two sensors in the same cell. To remedy this problem, reduce the Cell Transmission Model Time to reduce the cell lengths and prevent this from occurring.
- **Status: Ineligible Date** → A date or dates in the 'Analysis Dates' box do not match any available file. Check that the dates are correct, the Working Directory in the 'Input Parameters' box is set to the correct working directory, and that these files are in a folder labeled 'Loop Detector Data' in this working directory.
- **Status: Out of Memory** → The memory of the computer has been exceeded. Try reducing the size of the problem, such as decreasing corridor length or the number of dates being analyzed, or increase the Cell Transmission Model Time.

The Model's End Result

When the program is finally complete, four useful pieces of data are presented for the user. Each of these pieces is discussed below:

Degree of Optimality

In the 'Overall Progress' box, a completed result will be given as a percentage of optimality. This percentage represents the benefit captured, defined by the value of benefit in Liu and Danczyk (2009), relative to the benefit captured for a complete deployment. Thus, if it reveals that the result is 80 percent optimal, then the benefit captured could, say, be 800 while a deployment at all sites would have captured 1000. This is assistive in gauging the benefits of allowing an extra sensor to be budgeted.

Overall Progress:	Task Complete
Status:	Standing By
Result:	84.52% Optimal

Figure 11: Overall Progress Box when Program is Complete

Proposed Configuration

Using the model, the program calculated recommended sensor allocations for the times defined in each of the provided dates. Since the allocations may differ for different days, the program tallied the frequency for each site where an allocation was recommended. After checking all dates, recommendations were made for each given site depending on this frequency and the beneficial strength assessed by the model. Each individual cell location can be determined in the 'New Configuration' box under the 'Location' column. These locations are given as stations in lengths of feet from the start point of the corridor. The frequency that sensors were allocated to these sites across all dates is given under the 'Sensor Allocation Frequency' column. The final recommendations are given under the 'Recommended Allocation' column, where 1 is a suggested site for allocation and 0 is a site not suggested for allocation.

View Allocation | Clear New Sensors

NEW CONFIGURATION

Location (feet)	Sensor Allocation Frequency	Recommended Allocation
0	0.00%	0
441	0.00%	0
882	100.00%	1
1323	100.00%	1
1764	0.00%	0
2205	0.00%	0
2646	0.00%	0
3087	0.00%	0
3528	0.00%	0
3969	100.00%	1
4410	0.00%	0
4851	0.00%	0
5292	0.00%	0
5733	0.00%	0
6174	0.00%	0
6615	0.00%	0
7056	0.00%	0
7497	0.00%	0
7938	100.00%	1
8379	100.00%	1
8820	0.00%	0
9261	100.00%	1
9702	100.00%	1
10143	0.00%	0
10584	100.00%	1
11025	0.00%	0
11466	0.00%	0

Figure 12: New Configuration Box

Repair Priority

Using the model, the program estimated the priority for repair for the recommended deployment, including any existing sensors. This priority for repair would assist users faced with a deployment suffering from sensor failures and only a limited budget for repair. Priority was determined by benefit loss that would occur given the removal of that sensor. Thus, the sensor location with the highest priority (priority 1) would create the greatest loss if removed.

Under the ‘Repair Importance’ box, the priority for repair is listed. The specific sensor is defined by the site in which it is located, shown beneath the ‘Location’ column which represents the stationing from the corridor’s beginning, in feet. In the ‘Priority’ column, the ranked priority is given, 1 being of the highest priority.

Clear Repair

REPAIR IMPORTANCE	
Location (feet)	Priority
37044	1
33075	2
32634	3
9702	4
37485	5
9261	6
28224	7
23814	8
23373	9
29988	10
882	11
11907	12
30429	13
1323	14
22932	15
27783	16
12348	17
22491	18
7938	19
16317	20
15876	21
8379	22
20727	23
24696	24
27342	25
33516	26
34398	27
3969	28
10584	29
18081	30
28665	31
37926	32

Figure 13: Repair Priority Box

Performance Measures

With the given configuration recommendations and knowledge of traffic flows over the corridor, the program can calculate both the ground truth state of various performance measures as well as the assessment that would be made by the given sensor configuration. Three performance measures are studied: Vehicle-Hours Traveled (VHT), Vehicle-Miles Traveled (VMT), and Bottleneck Activity Time (BAT). The first two are commonly-used performance measures applied by transportation agencies around the United States and world. The third is a performance measured defined and used in Liu and Danczyk (2009).

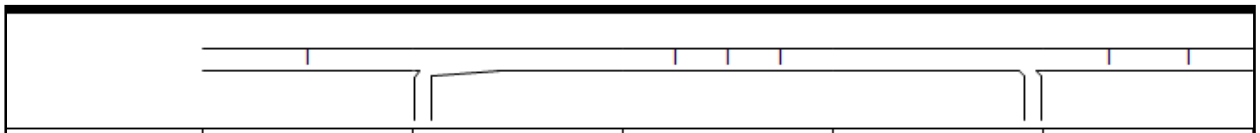
PERFORMANCE MEASURES			
	Measured	Ground Truth	Rel. Error
VHT	4139.20	4021.98	2.91%
VMT	122009.42	121956.82	4.31E-04
BAT	15.93	15.93	0.00%

Figure 14: Performance Measures Box

In the 'Performance Measures' box, these three performance measures are computed. Under the 'Measured' column, the generated value is what the recommended sensor configuration would produce if deployed. Under the 'Ground Truth' column, the generated value is what the ground truth state is. The 'Rel. Error' column reveals the relative error between the measured assessment and the ground truth state.

Mapping Feature

The program includes a feature that allows a graphical representation of the freeway corridor to be generated. This corridor is designed by the parameters entered in the 'Freeway Characteristics' box on the main page and the sensor allocation determined by the model. While not to scale, the intent is to give the user a basic understanding of where the sensors have been deployed relative to on-ramps or off-ramps.



Type:	Pipeline	Entrance Ramp Lane Addition	Pipeline	Exit Ramp Diverge Area	Pipeline
Length:	1056 feet	100 feet	4652 feet	100 feet	2270 feet
Lanes:	3	3	4	4	4
Capacity:	6900 veh/hr	6900 veh/hr	9200 veh/hr	9200 veh/hr	9200 veh/hr

Figure 15: Graphical Representation of Freeway Characteristics and Sensor Locations

Once a sensor configuration has been determined through the model, a map with this sensor deployment can be generated. The best and easiest method is to use the button on the main page above the 'New Configuration' box that is labeled 'View Allocation'. Pressing this button will switch to the 'Map' tab and generate the graphical representation of the freeway.

Another method would be to manually switch to the 'Map' tab. On this tab, three new buttons can be found. Their functions are listed:

- **Create Map**: Generates the graphical map and identifies segment type, length, number of lanes, and capacity. This data is based on the characteristics defined in the 'Freeway Characteristics' box on the 'Input Parameters' tab.
- **Map Sensor Locations**: Generates the graphical map and identifies segment characteristics, as well as places sensors. The sensor data is based on the results generated by the model. It is important to verify that the model's result is up-to-date.
- **Clear Map**: Clears the graphical map and results.

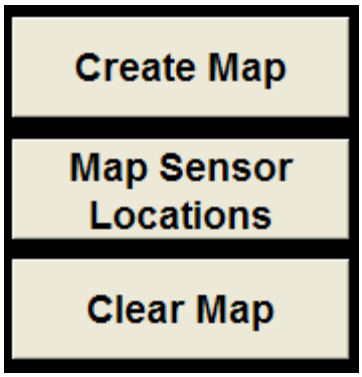


Figure 16: Buttons on the 'Map' Tab

When the ‘Create Map’ or ‘Map Sensor Locations’ button is pressed and the process is completed, a green indicator should appear near the ‘Status’ box to identify that the task is complete. Otherwise, a red indicator will appear with an input error message. Generally, fixing this problem requires ensuring that the ‘Freeway Characteristics’ are correct and that the model has computed sensors to deploy.

Stored Data

A third tab is provided in this document, labeled as ‘Stored Data’. This tab does not play any role in the program and is intended as a storage site for date if multiple runs or freeway sets are desired.

Loop Detector Data Format

The individual loop detector data files follow a very specific format so that they can be accurately interpreted by the tool. The detector identification (or Flow ID) serves as column headers while time interval serves as the row headers. Unlike the rest of the tool, a space is necessary between the headers and the data set in order to function properly. The data collected in this file is the counts at any given time for a specified time interval and detector site. The figure below illustrates how the .csv document should appear.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Detector ID:	1	2	3	4	5	6	7	8	9	10	
2	Time:												
3	2:00:30 PM		24	9	3	2	15	11	2	2	2	7	
4	2:01 PM		28	12	8	3	14	10	4	3	0	0	
5	2:01:30 PM		23	5	4	2	3	7	2	1	3	3	
6	2:02 PM		21	9	7	6	8	9	3	2	2	3	
7	2:02:30 PM		16	7	6	12	12	9	3	4	5	9	
8	2:03 PM		25	8	1	20	10	4	5	3	4	1	
9	2:03:30 PM		21	6	4	3	12	10	7	1	2	6	
10	2:04 PM		18	6	5	9	7	4	2	4	1	2	
11	2:04:30 PM		22	2	4	11	11	11	0	1	2	2	
12	2:05 PM		17	8	4	1	14	8	1	2	2	4	
13	2:05:30 PM		15	6	3	3	8	8	3	4	2	2	
14	2:06 PM		30	7	4	5	10	5	0	3	3	6	
15	2:06:30 PM		24	7	3	1	8	8	1	6	5	5	
16	2:07 PM		22	7	2	2	2	13	0	5	3	4	
17	2:07:30 PM		30	9	7	0	7	7	2	3	3	1	
18	2:08 PM		16	10	4	21	6	6	2	5	2	0	
19	2:08:30 PM		25	13	4	11	15	10	3	1	3	5	
20	2:09 PM		23	3	4	0	17	8	3	3	2	3	
21	2:09:30 PM		23	8	3	4	9	5	2	0	3	5	

Figure 17: Loop Detector Data File

Acknowledgements

This tool was developed by Henry X. Liu and Adam Danczyk at the University of Minnesota in regards to research surrounding optimized sensor deployments around freeway bottlenecks. Financial support for conducting this research was provided by California Department of Transportation (Caltrans) and the University of California, Berkeley. Sincere thanks go to Mr. JD Margulici and Dr. Jeff Ban of UC-Berkeley for their contributions and collaboration in this project.

References

Daganzo, C.F. (1994), The Cell Transmission Model: A Dynamic Representation of Highway Traffic Consistent with the Hydrodynamic Theory, *Transportation Research Part B*, Volume 28B, No. 4, pp. 269-287.

Daganzo, C.F. (1995), The Cell Transmission Model, Part II: Network Traffic, *Transportation Research Part B, Volume 29B*, No. 2, pp. 79-93.

Liu, H.X. and Danczyk, A. (2009), Optimal Sensor Locations for Freeway Bottleneck Identification, submitted to *Computer-Aided Civil and Infrastructure Engineering*.

Rural Issues with Optimal Sensor Placement for Transportation Applications

by

Patrick McGowen
Assistant Professor

Western Transportation Institute
College of Engineering
Montana State University

A report prepared for

J. D. Margulici
California Center for Innovative Transportation
University of California–Berkeley
2105 Bancroft Way, Suite 300
Berkeley, CA 94720-3830

For support of
California PATH Project
TS-603 – Optimal Sensor Requirements

August, 2008

DISCLAIMER

The opinions, findings and conclusions expressed in this report are those of the author and not necessarily those of the California Department of Transportation, the California Center for Innovative Transportation or Montana State University.

ACKNOWLEDGEMENTS

Thanks to the students who helped gather information for this research and assisted in the writing of this report, including Lili Liang and Scott Randall. Also, thanks to Chris Strong and Doug Galarus for their guidance and support. Thanks especially to CCIT for including WTI in this project.

TABLE OF CONTENTS

List of Tables

1. INTRODUCTION

Many types of sensors are used in managing transportation systems. These sensors supply critical information used by transportation managers for a variety of purposes such as real-time response to changes in travel and traffic conditions, or planning for improvements to the transportation system. In the absence of well developed methodologies to plan the deployment of these sensors, the processes that are used in selection of their location do not always follow a set of criteria that optimize their usefulness. Development of location selection guidelines will assist transportation managers in making the most efficient use of these sensors. This is particularly true in rural areas, given the unique challenges related to topography and remoteness. This report provides an overview of some of the issues and concerns encountered in locating sensors in rural areas that are used for assessment of travel conditions, incident detection, incident verification and collection of planning data. The sensors discussed include road weather information system (RWIS) stations, closed circuit television (CCTV) cameras and inductive loop detectors. Dynamic message signs (DMSs) are also examined as they are commonly co-located with other devices.

2. APPLICATIONS

Intelligent transportation systems (ITS) employ sensors for a number of purposes. Sensor applications discussed in this report include travel time estimation, incident detection, incident verification, and collection of planning data. For more information on the applications discussed in this report and other rural ITS applications, refer to the Rural ITS Toolbox (http://www.itsdocs.fhwa.dot.gov/jpodocs/repts_te/13477.html).

2.1. Travel Time Estimation

ITS elements are useful in determining travel times along individual routes and/or highway networks. Real-time measurements of travel times can help in determining problem areas within the transportation system and provide valuable traveler information. Determining the optimal sensor placement for travel time estimation is a key issue in urban areas where congestions is a major factor affecting these estimates. Travel times in rural areas are less impacted by traffic congestion than by incidents caused by weather, crashes and landslides. Therefore, the focus of placing sensors in rural areas should be on detecting these incidents, or their causes, which is discussed in section 2.2.

If travel times are to be estimated in rural areas, use of permanent sensors (such as inductive loop detectors) following methodologies employed in more urban environments will result in a large number of sensors to cover the long distances involved. The use of probe vehicles is a more feasible approach. Many trucking companies utilize a service that will track their trucks in real-time through a satellite connection to in-vehicle global positioning systems. The American Transportation Research Institute (2005) showed that real-time corridor travel times could be estimated based on truck data already collected by these services.

For these reasons travel time estimation using permanent sensors will not be discussed further in this report.

2.2. Incident Detection and Verification

Unplanned incidents and events in rural areas can include crashes, weather events and landslides. If cellular coverage exists, the incident is discovered and reported quickly by motorists with cell phones. In areas without cellular coverage call boxes can be used.

CCTV cameras can be helpful for incident and event verification, but the number of cameras required to provide visual coverage of the entire rural network is cost prohibitive. CCTV cameras placed in rural areas should have a good vantage point that provides views of a large area. Camera effectiveness is also optimized by placing them in areas where incidents and events are commonly known to occur.

Most efforts toward incident detection relate to weather events. With RWIS, weather events can be detected immediately and even predicted with an appropriate weather model.

2.3. Planning Data

Another use of sensors in rural areas is the collection of planning data, primarily consisting of traffic counts for estimation of average annual daily traffic. This is typically accomplished with inductive loop detectors and portable road tubes.

3. TYPES OF SENSORS

The sensors discussed in this chapter are related to the applications discussed in Chapter 2 (Table 1). Dynamic message signs, which are not a sensor, are also discussed because sensors and DMSs are often located together in rural areas due to power and communication issues. For a discussion of co-locating DMSs with sensors, refer to Section 5.5.

Table 1: Matrix of Sensors and Applications

	Application		
Sensor	Incident Detection	Incident Verification	Planning Data
RWIS	O	O	
CCTV		O	
Loops			O

Although inductive loops are commonly used for incident detection in urban areas, they are impractical in rural areas due to the long distances and lower traffic volumes. In rural areas, loops are used almost exclusively for collecting planning data. Weather events can be detected and even predicted by RWIS stations. CCTV is often used to verify events.

3.1. Road Weather Information Systems

Having information about the road surface is especially important when dealing with adverse weather conditions. Road weather information systems (RWIS) stations are used to collect area weather information as well as road surface conditions. Information collected includes wind speed, air temperature, surface temperature, precipitation amount and type, visibility, and road condition.

RWIS data can be sent to traffic management centers, consolidated to an Internet webpage, or used as inputs to local weather or pavement surface conditions prediction models. State DOTs can use this data to plan winter maintenance activities (plowing and deicing) and to detect weather events. RWIS stations are commonly located in areas where inclement and abrupt changes in weather often occurs such as mountain passes. RWIS stations are more effective when there is a network of stations with wide coverage of the rural region.

3.2. Closed Circuit Television

Closed circuit television (CCTV) camera systems include a network of cameras that collect images of current roadway conditions. The images are sent to a monitoring location where they can be viewed and analyzed. These images can provide information about area traffic and roadway conditions, be used to detect and verify reported traffic incidents, and assist in activity management.

There are many benefits of being able to monitor current roadway conditions at important locations. Information about the current status of the roadway can aid in making the area safer for travel. CCTV can be used to detect incidents and help with incident response time. CCTV can provide information about roadway surface conditions, which is important for winter maintenance and for traveler information. Images from CCTV cameras can be made accessible to the public via the Internet, which can help ensure that drivers are prepared for hazardous

conditions. CCTV cameras can also be a useful tool in providing security in locations such as rest areas. Some state DOTs use CCTV cameras to view DMSs to verify the current message being posted.

3.3. Inductive Loop Detectors

Inductive loop detectors (also referred to as loop detectors) are placed underneath the road surface and detect vehicles that pass over them. Information provided by loop detectors can be analyzed to determine patterns and volumes over time, or the detectors can be used to provide useful data in real-time. In urban areas real-time information collected by loop detectors relates to incident detection, traffic monitoring, or travel-time forecasting. Due to the traffic and travel challenges in rural areas, they are generally used solely for planning or analysis purposes such as:

- Estimates of average annual daily traffic (AADT) for roadway segments;
- Pavement design data such as lane distribution, directional distribution, percent trucks and possibly truck weights; and
- Operation characteristics such as peak-hour flow and average speed.

This report focuses on the methods used to determine optimal locations for collecting statewide AADT data. To accomplish this, the sampling method relies upon permanent or continuous count locations (comprised of inductive loops) and portable count locations using road tubes.

It should be noted that there are numerous alternatives to inductive loops such as magnometers, piezoelectric sensors, and a wide range of non-intrusive detectors that use radar, sound, video image processing and other technologies to detect vehicles. For more information on vehicle detection technologies visit the Vehicle Detector Clearinghouse web site at New Mexico State University (http://www.nmsu.edu/~traffic/).

3.4. Dynamic Message Signs

Dynamic Message Signs (DMSs) can be used to display real-time information to drivers about hazardous surface conditions, traffic problems, road construction, or any other issues that may affect traveler safety or convenience. Alerting drivers of upcoming conditions may result in safer driving actions ranging from more attentiveness, to reducing speed, to stopping to put chains on the vehicle. At a broader level, DMS can be used to influence motorist route selection when an incident occurs. Thus they are often placed upstream of major junctions or interchanges.

The signs can be fixed or portable depending on whether the need is temporary or permanent. For example, portable signs work well in areas that are under construction. Fixed signs are ideal in locations where the drivers' need for safety information is continual, such as road segments where traffic incidents or weather events are a common occurrence. This document focuses on locations for permanent DMS installations.

4. RURAL CHALLENGES

Rural areas are, by nature, more remote and sparsely populated than urban areas. In California, rural regions are often characterized by rugged or inhospitable terrain such as mountains and deserts. These elements provide challenges that affect sensor location selection, described below:

- There is limited access to power. It can be costly to run electricity from the nearest access point to power the sensors. Other power sources such as solar panels and batteries can be employed, but they require more components on the roadside that require maintenance and are susceptible to vandalism. These practical concerns with stand alone power systems should be considered when selecting sensor locations.
- Communications to allow remote data retrieval and incident notification are critical to making use of sensors. Ideally sensor locations should have access to landline communication. Cellular networks offer another option. If cellular is used, bandwidth and connection issues should be considered, as some system designs will lead to excessive cellular charges.
- Maintenance of sensors requires trained staff. Rural districts face staffing issues associated with ITS because staff with appropriate training to calibrate and maintain ITS devices are often stationed at the district office which may be some distance from a remote ITS device. This affects location selection because it increases travel costs associated with checking and maintaining sensors. If possible, locations should be selected that are readily accessible to maintenance staff. Sensor locations should also have safe, convenient access, such as wide shoulders or pullouts. If sensors are to be installed in more remote locations, consideration should be given to more robust (and typically more expensive) installations that will perform reliably of longer maintenance intervals.
- Rural areas have few alternate routes. If motorists are to choose an alternate route based on information from an incident detection system, they may need to be notified many miles upstream at the nearest junction. Furthermore, they may still choose to continue on their original route, as the travel time added by taking the alternate route could offset the benefit of avoiding the incident.
- Rural areas have longer emergency response times. Areas with long emergency response times and high crash frequency may be ideal locations for CCTV cameras.

5. LOCATION SELECTION

The following sections discuss location selection criteria that are specific to each type of sensor mentioned—RWIS, CCTV, inductive loops, and DMS. The last section provides guidance on location selection criteria that are common to all devices. These criteria are summarized in Table 5 in Chapter 6. In light of the variability in their intended function and the uniqueness of each deployment site, the guidance on sensor placement, offered in this report, is intended more to highlight rural issues, rather than to provide implementation and design guidelines. For an example of applying these principles of location selection in a specific region, refer to Strong et al. (2005).

5.1. Road Weather Information Systems

RWIS stations generally are most useful in areas of adverse weather conditions. These could be areas with high variability in weather patterns or areas that are prone to weather events that affect the transportation system such as high winds, fog, precipitation and freezing temperatures. Considerations for siting RWIS stations include locations with:

- Known weather issues such as mountain passes (ice and snow), bridge decks (ice), and valleys and shaded areas (icy patches)
- Frequent traffic during adverse weather conditions, such as ski areas
- High amounts of snow, rain, fog, or wind
- A higher frequency of weather-related accidents
- More likelihood of flooding (FEMA classification “A”)
- A higher frequency of road closures caused by storms, avalanches, or weather-related crashes.

Although it is helpful to place RWIS stations in areas with weather problems for incident detection at specific locations, they can also be used collectively to monitor regional weather and pavement surface conditions. Local prediction models can also be developed that utilize regional weather forecasts and a system of several local RWIS stations. For regional monitoring and prediction RWIS stations should be located for the best area coverage and spread across the region, with more locations in areas of higher variability. Ballard et al. (2002) recommends that a licensed meteorologist provide guidance on RWIS station placement. For more information on using RWIS for regional models and coordinating weather data with other agencies, refer to the Federal Meteorological Handbook (http://www.ofcm.gov/fmh-1/fmh1.htm). More detail on the generalized location guidelines summarized above for siting RWIS stations can be found in several publications:

- RWIS Environmental Sensor Station Siting Guidelines (http://ops.fhwa.dot.gov/publications/ess05/ess05.pdf)
- RWIS Volume 1: Research Report (http://onlinepubs.trb.org/Onlinepubs/shrp/SHRP-H-350.pdf)
- RWIS Volume 2: Implementation Guide (http://onlinepubs.trb.org/Onlinepubs/shrp/SHRP-H-351.pdf)

5.2. Closed Circuit Television

CCTV is used to provide a visual picture of current conditions at remote locations. While other sensors provide statistical measures of current conditions, CCTV can be used to visually confirm and monitor conditions. Criteria used to select RWIS station locations also apply to locating CCTV cameras when they are to be used to verify weather events.

It is common to use CCTV cameras at intersections, junctions and major interchanges to monitor traffic. Because of merging traffic, these locations are susceptible to congestion and higher crash frequencies. Monitoring the traffic at these locations can be beneficial because it allows for quicker deployment of emergency vehicles in case of an accident. Monitoring video of how an intersection performs during certain traffic events, such as closing time for a ski area, can be useful in determining ways to alleviate the problem.

Common crash locations, whether at intersections or other road segments, should be considered potential sites for CCTV cameras.

CCTV is also commonly used for security purposes. Places where people congregate or where vehicles are left unattended, such as rest areas or parking areas, can have security problems. It may be desirable to install CCTV cameras in these locations for the security they can provide.

CCTV can be used to verify that a nearby DMS is working and displaying the appropriate message. The goal of verifying DMS messages by itself probably does not justify a CCTV camera installation, but is an additional benefit if other location selection criteria are met.

For CCTV to be effective, the camera needs a clear unobstructed view of the area being monitored. Using the cameras on straight, open stretches of road optimizes the range of effectiveness of the camera. To ensure that the best picture quality is achieved, the camera should be mounted to a fixed, steady structure such as a bridge or overpass. Cameras mounted on poles can be affected by wind and provide poor quality images. If the camera is mounted on a structure that vibrates with passing traffic, the picture may look blurry and the images may be of little use. Locations to consider for siting CCTV cameras include:

- Location of existing or planned RWIS stations (see previous section)
- Locations that provide a view of existing or planned DMS
- Locations with high frequency of crashes
- Rest areas
- Major intersections or interchanges
- Structures such as overpasses
- Locations with a clear view

Washington and Wisconsin use one-mile spacing between CCTV cameras in urban areas as a general rule (Strong et al., 2005). This level of CCTV camera density is not feasible in rural areas. Typically the number of CCTV cameras deployed is based on budget constraints, with locations limited to the highest priority areas.

5.3. Loop Detectors and Planning Data

Inductive loop detectors can be used for incident detection and ramp metering in urban areas, but in rural areas are typically only used for collecting planning data. Inductive loops are used for permanent count stations at locations that are chosen based on a pre-determined sampling scheme. These permanent count stations are supplemented by short-term counts made with portable road-tube counters. The layout of the permanent count stations (number and location) is determined to minimize the potential error in traffic estimates across the region or state. More detail on the general approach, described below, can be found in the FHWA Traffic Monitoring Guide (http://www.fhwa.dot.gov/ohim/tmguide/). This approach, although used, is not specific to rural areas.

First, roadways are categorized by facility type as shown in Table 2. Generally, the seasonal and weekly variations in traffic are similar for roads of the same facility types.

Table 2: HPMS Functional Code (FHWA, 2001)

Rural Functional System Codes	Urban Functional System Codes
1 Principal Arterial Interstate	11 Principal Arterial Interstate
2 Other Principal Arterial	12 Principal Arterial Other Fwys & Exp
6 Minor Arterial	14 Other Principal Arterial
7 Major Collector	16 Minor Arterial
8 Minor Collector	17 Collector
9 Local	19 Local

Next, these functional classifications are combined into functional groups with similar seasonal and weekly variations in traffic. FHWA (2001) recommends three to six groups, or more if needed to account for regional differences. A potential functional group categorization is shown in Table 3.

Table 3: Minimum Groups (FHWA, 2001)

Road Function	HPMS Functional Code
Interstate Rural	1
Other Rural	2, 6, 7, 8
Interstate Urban	11
Other Urban	12, 14, 16, 17
Recreational	Any

The permanent counters are assigned within these groups. The minimum number of permanent count stations needed within a functional group depends on the variability of daily traffic counts within that group. More permanent count stations should be placed within groups of higher variability. The following equation can be used to determine the minimum number (n) of count stations for a desired precision (D*) and a variability of the facility group (in this case the coefficient of variation, C).

$$D^* = t_{1-d/2,n-1} * \frac{C}{\sqrt{n}}$$

Minimum count stations for groups with different coefficients of variation have been calculated and listed in Appendix A. The coefficient of variation should be based on previous counts. If the coefficient of variation is unknown, an estimate can be chosen from the ranges in Table 4.

Table 4: Variation Coefficient by Area Function (FHWA, 2001)

Area Function	Variation Coefficient
Urban Area	<10%
Rural Area	10%~25%
Recreational Area	>25%

If resources exist for additional count stations above the minimum for each functional group, they should be assigned proportionally to these minimum numbers. More count stations will yield more accurate results. The optimum number, above the minimum, is open for debate. As a point of reference, in Iowa, with 8,909 miles of state roads, Souleyrette and Pattnaik (2003) found that the 130 permanent detectors yielded adequate results. Resources for collecting planning data should be appropriated such that there are 10 to 20 portable road-tube counters for every permanent count station (Ross et al., 2004).

Thus new permanent count stations should be located on roadways within a functional group such that, when compared to other groups, the proportion of counters located on the types of roadways within that group is equivalent to the minimum number, when compared to the minimum number of other groups. Ideally, locations on roadways within these functional groups should be selected randomly. However, to improve accuracy, they should be located on straight segments of roadway.

5.4. Dynamic Message Signs

DMSs are commonly used in areas where there is a need to detour traffic or warn motorists of downstream conditions. A typical use for DMSs in rural areas is to display information pertaining to weather conditions, such as warning drivers of icy conditions as they approach mountain passes, and whether snow chains are needed. Such a sign would usually be placed at the bottom of the pass or in advance of a junction, turn-around point, or a pull-out/chain-up area.

The signs must be large enough that the motorist can read them at highway speeds. Thus, they must be placed in locations where they can be seen from an adequate distance such as a straight stretch of roadway longer than 800 feet. The signs must also be visible at all times of the day, giving consideration to sun glare or headlight reflection. Considering DMS use in rural areas, they should be placed:

- Two miles prior to major junctions
- Two miles prior to snow chain areas
- After 800 feet of straight road
- Locations that meet placement requirements of guidance signing in the Manual on Uniform Traffic Control Devices
- Upstream of common weather events

5.5. All Devices

There are several location criteria that are common to any ITS device in the rural environment, such as availability of power and communications. Although solar power and satellite communication can be used, landlines for power and communication are preferred, as previously discussed.

Sensor maintenance costs are potentially of greater concern in rural environments, as travel distances, and associated travel costs, to perform inspection and maintenance activities can be significantly higher. Travel time (i.e., the distance from the sensor to the maintenance office or district office) is an important consideration. If a remote area has significant challenges that could benefit greatly from ITS elements, travel costs should not preclude the placement of an ITS device, but they need to be considered.

It is often more cost effective to install new ITS elements in conjunction with new construction projects. When ITS elements are included in construction projects, economies of scale can be realized with design, mobilization and traffic control. It is also easier to run power and communication lines to ITS elements while the road is under construction. Disruption to traffic during the installation is also reduced if the ITS elements are installed during construction since the traffic would already be disrupted.

The proximity of each element to other ITS elements should be considered. Similar types of devices should not be duplicated in the same area. Similarly, if there are large gaps in the system, it may be desirable to install an element where data can be obtained for the underserved area. Note that if CCTV cameras without pan/tilt capabilities are used, two may be used in one location to view each direction.

Because of the power, communication and maintenance challenges in rural areas, it may be advisable to take a node approach to ITS devices. A single communication and power hub may be set up for CCTV, RWIS, loops, and DMS in a single area. The devices could be mounted on the same structure, or within the same area as long as they are close enough together to share the communication hub.

Regardless of the sensor type, the following location considerations should be made:

- Planned construction project
- Available power
- Available communication
- Close to maintenance yard (e.g., within three hour drive)
- Directly adjacent to maintenance yards
- No existing device of the same type nearby (e.g., within two miles of similar device)
- Co-locate ITS devices
- Good access (e.g., near a pullout)

6. SUMMARY

This report summarizes the primary attributes of good locations for sensor placement in rural areas. To optimize sensor placement, locations should be prioritized for a district or region. First, location attributes that are important for the district should be identified from those in Table 5. Note that it may be beneficial to place DMS upstream of locations with many of the criteria listed (e.g., mountain passes and common weather events) placing the DMS at the problem locations may not be beneficial. Second, a set of ranking criteria or point system should be developed for each of these criteria. An example that was developed by Strong et al. (2005) for prioritizing DMS and CCTV cameras is presented in Appendix B. Third, the ranking criteria should be applied to roadway data to develop a prioritized list of locations.

Table 5: Location Selection Criteria

Location Criteria	RWIS	CCTV	Loops	DMS
Mountain passes	X	X		
Ski areas	X	X		
High wind, rain, snow, fog	X	X		
Common icy conditions	X	X		
Shaded areas	X	X		
Bridges	X	X		
High proportion of weather related crashes	X	X		
Flooding locations	X	X		
High frequency of road closures	X	X		
View of DMS		X		
High frequency of crashes		X		
Rest areas		X		
Major intersections and interchanges		X		
Structures		X		
Good view		X		
Straight road		X	X	X
Upstream of junctions				X
Upstream of chain up areas				X
MUTCD				X
Upstream of common weather events				X
Planned construction project	X	X	X	X
Available power	X	X	X	X
Available communication	X	X	X	X
Can visit from maintenance yard in one day	X	X	X	X
At maintenance yard	X	X	X	X
Not within 2 miles of same type of device	X	X	X	X
Co-located with other devices	X	X	X	X
Good access	X	X	X	X

7. REFERENCES

American Transportation Research Institute. 2005. *Methods of Travel Time Measurement in Freight-Significant Corridors.* Submitted to Transportation Research Board. January 2005.

Ballard, L., A. Beddoe, J. Ball, E. Eidswick, and K. Rutz. 2002. *Assess Caltrans Road Weather Information Systems Devices and Related Sensors.* Prepared for Caltrans New Technology and Research Program. July 2002.

Boselly, S. E., Doore, S., Ernst, D. *Road Weather Information Systems Volume 2: Implementation Guide.* Strategic Highway Research Program, 1993.

Boselly, S. E., Doore, S., Thornes, J., Ulberg, C., Ernst, D. *Road Weather Information Systems Volume 1: Research Report.* Strategic Highway Research Program, 1993.

Manfredi, J., T. Walters, G. Wilke, L. Osborne, R. Hart, T. Incrocci, T. Schmitt. *Road Weather Information System Environmental Sensor Station Siting Guidelines.* Federal Highway Administration Report No. FHWA-HOP-05-026. 2005.

Ross, R., E. Prassas, and W. Mcshane. 2004. *Traffic Engineering*, third edition. Pearson Education Inc.

SAIC, Castle Rock Consultants, Transcore, WTI. *Rural ITS Toolbox*. [web document] http://www.itsdocs.fhwa.dot.gov/jpodocs/repts_te/13477.html.

Souleyrette, R. and S. Pattnaik. 2003. *Designing a traffic monitoring program using land use change detection.* Center for Transportation Research and Education, Iowa State University.

Sreedevi, I. and J. Black. 2001. *Loop Detectors*, Final Report [web document] http://www.calccit.org/itsdecision/serv_and_tech/Incident_management/Incident_detection/loop_detectors/loop_detectors_report.html.

Strong, C., S. Torgerson, and B. Snyder. 2005. *Development of Criteria to Identify Locations for ITS Deployment*: Final Technical Report. Prepared for USDOT RITA and Oregon DOT. June 2005.

USDOT. 2001. *Traffic Monitoring Guide*. Federal Highway Administration. May 2001.

USDOT. 1997. *Field Test of Monitoring of Urban Vehicle Operations Using Non-Intrusive Technologies.* May 1997.

8. APPENDIX A: ESTIMATED NUMBER OF COUNT STATIONS

The following table provides the number of permanent count stations (N) needed for a 10 percent precision interval for a functional group with a given coefficient of variation.

Table 6: Number of Count Stations (N) Based on Variability of Traffic (C)

Area Function	C	N
urban	1%	2
	2%	3
	3%	3
	4%	3
	5%	3
	6%	4
	7%	4
	8%	5
	9%	5
Rural	10%	6
	11%	7
	12%	8
	13%	8
	14%	10
	15%	11
	16%	12
	17%	14
	18%	15
	19%	16
	20%	18
	21%	19
	22%	21
	23%	23
	24%	25
	25%	26
Recreational	26%	28
	>27%	30

9. APPENDIX B: TABLES TAKEN FROM STRONG ET AL. 2005

Table 7: DMS Location Criteria

	Positive Criteria	+ 2 pts	+ 1 pt
	Incident Prevention		
1.	Percent of crashes attributable to weather	> 50%	20 to 50%
2a.	Presence of sharp horizontal curvature		10 second duration curve with radius tighter than 75 percent of recommended radius at e=0.04
2b.	Presence of sharp vertical grade		1 mi. with avg. grade of >5%
	Incident Management		
3.	Crash rate compared to state mean crash rate for similar highway segments*	>2 σ higher	1-2 σ higher
4.	Vehicle-hours of delay for road closures*	>100,000	10,000 to 100,000
5.	Vehicle-hours of delay for incidents*	>200,000	20,000 to 200,000
6.	Average spacing between state highway intersections	>40 miles, and <4 mi. to nearest intersection (Rural)	20 to 40 miles, and <4 mi. to nearest intersection (Rural)
		>10 miles, and <2 mi. to nearest intersection (Urban)	5 to 10 miles, and <2 mi. to nearest intersection (Urban)
7.	Product of average interchange or access point spacing and mainline traffic volume	>500,000	200,000 to 500,000
8.	Ratio of ramp to mainline volume	>0.5 (Rural)	0.2 to 0.5 (Rural)
		>0.3 (Urban)	0.15 to 0.3 (Urban)
9.	Proximity to freeway-to-freeway interchange		<2 mi.
10.	Percentage of truck traffic	>35%	22 to 35%
	Non-Incident Congestion Management		
11.	Percent of vehicles in congestion	>75%	50 to 75%
12.	Annual average daily traffic	> 50,000	20,000 to 50,000
13.	Total visitation of attractions within five miles		> 1 million per year
	Weather Warnings		
14.	High wind areas – using wind power value	6 or 7 (>17.9 mph)	5 (16.8 – 17.9 mph)
15.	Located in area susceptible to floods		"A" FEMA classification
16.	Proximity to RWIS		< 10 mi.
	Enabling Criteria		
17.	Distance from maintenance yard		> 50 mi.

	Negative Criteria	– 4 pts	– 2 pts
1.	Distance to nearest DMS	< 2 mi.	2-5 miles
2.	Travel time from regional office		> 3 hours (-1 pt)

* - Over three-year period

Table 8: CCTV Location Criteria

	Positive Criteria	+ 2 pts	+ 1 pt
	Incident Detection		
1.	Crash rate compared to state mean crash rate for similar highway segments*	>2 σ higher	1-2 σ higher
2.	Proximity to freeway-to-freeway interchange	1 mile	2 miles
3.	Location of nearest major interchange (urban) – ramp to mainline volume ratio of 0.15 or greater	1 mile	2 miles
4.	Proximity to bridge or tunnel	In segment	
	Incident Response and Management		
5.	Location of nearest camera (urban)	>2 miles	1-2 miles
	Non-Incident Congestion Management		
6.	Percent of vehicles in congestion	>75%	50 to 75%
	Pre-Trip Traveler Information		
7.	Proximity to mountain pass	<1 mile	1-4 miles
8.	Proximity to major attraction	<1 mile	1-2 miles
9.	Proximity to ski area	<1 mile	1-2 miles
	Maintenance		
10.	Location of nearest maintenance yard	>30 miles	20-30 miles
11.	Location of nearest current and proposed RWIS	<1 mile	1-2 miles
	Security and Verification		
12.	On roads entering state, facing inbound traffic	>50,000 AADT	>10,000 AADT
13.	Location relative to DMS	<2 miles	2-5 miles

	Negative Criteria	– 4 pts	– 2 pts
1.	Distance to nearest CCTV	<1 mile	1-2 miles
2.	Travel time from regional office		> 3 hours (-1 pt)

* Over three-year period