PARTNERS FOR ADVANCED TRANSPORTATION TECHNOLOGY

INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

HYBRID TRAFFIC DATA COLLECTION ROADMAP: PILOT PROCUREMENT OF THIRD-PARTY TRAFFIC DATA

FINAL REPORT FOR TASK ORDER 1

Alexandre M. Bayen, Ph.D., Principal Investigator
Mohammad (Ashkan) Sharafsaleh, P.E., Project Manager
Anthony D. Patire, Ph.D., Assistant Research Engineer

PATH Research Report
UCB-ITS-PRR-2013-1

Partners for Advanced Transportation Technology works with researchers, practitioners, and industry to implement transportation research and innovation, including products and services that improve the efficiency, safety, and security of the transportation system.

Partners for Advanced Transportation Technology
University of California, Berkeley
2105 Bancroft Way, Suite 300, Berkeley, CA 94720-3830
Phone: (510) 642-4522   Fax: (510) 642-0910   http://www.path.berkeley.edu

ISSN 1557-2269

OCTOBER 2013

HYBRID TRAFFIC DATA COLLECTION ROADMAP: PILOT PROCUREMENT OF THIRD-PARTY TRAFFIC DATA

FINAL REPORT FOR TASK ORDER 1

Prepared by:

Alexandre Bayen
Mohammad (Ashkan) Sharafsaleh
Anthony Patire
Joe Butler
Nate Bailey
Agathe Benoit
Sébastien Blandin
Joy Bonaguro
Raphaël Jatteau
Tristan Lall
Mark Merala
Ali Mortazavi
Boris Prodhomme
Samitha Samaranayake
Krute Singa
Nazy Sobhi
Fred Winik
Matt Wright
Yi Zhou

For:

CALIFORNIA DEPARTMENT OF TRANSPORTATION
DIVISION OF TRAFFIC OPERATIONS

ACKNOWLEDGMENTS

We would like to acknowledge the participation and support of the many people who, in their respective organizations, have contributed to the Hybrid Traffic Data Collection Roadmap project (Task Orders 1 and 2).

California Department of Transportation

This project would not have been possible without the interest, engagement, and financial support of our sponsors at Caltrans. Joan Sollenberger (Chief, Office of System Management Planning) and Nick Compin (Chief, Traffic Data and Performance Measures Branch) provided ongoing consultation and valuable input in an unfailing spirit of collaboration throughout the project. We are grateful for their confidence and involvement in the work.

We also received important assistance from the departments of Operations and Maintenance of Caltrans District 4 and District 8, including Sean Nozzari, Renato Dacquel, Renato Fortaleza, Thomas Ainsworth, Peter Acosta, and their respective groups. We thank them for their cooperation and support.

In addition, the Division of Research and Innovation and the Division of Operations, both at Caltrans Headquarters, and the Operations offices in District 4 and District 7 generously granted us interviews to explore the potential impact of purchasing probe data. We appreciate their time and insights.

Additional contributors

A variety of participants made significant contributions to the project, and we gratefully acknowledge their work. Thanks go to Nokia team members Jane Macfarlane, Andy Lewis, and Jeff Adachi; Iteris team members Karl Petty, Mark Merala, Andrew Moylan, Ricardo Reynoso, and Nick Hartman; and J.D. Margulici of Novavia Solutions.

We also received helpful input from the City of San Jose Transportation Department, the Contra Costa County Transportation Authority, the Metropolitan Transportation Commission, Berkeley Transportation Systems (a division of Iteris Inc.), and Delcan Corporation. Discussions with them expanded our understanding of current practices in the transportation community.

The PATH team

California PATH members at all levels and in all capacities provided the energy, dedication, and teamwork to develop and complete the project. We were fortunate to collaborate with such a talented group.

Management

We want to thank PATH Director Tom West and Program Manager Joe Butler, whose foresight and leadership made the project possible and propelled it forward. The work could not have proceeded without their efforts.

We would also like to express our appreciation to Samer Madanat, who served as original Principal Investigator for the project, and Ali Mortazavi for his contribution as original Project Manager.

Team members (staff, students, alumni, visitors, in alphabetical order)

Saneesh Apte (R&D Engineer), Nate Bailey (Undergraduate Research Apprentice), Agathe Benoit (Visiting Graduate Student Researcher), Sebastien Blandin (Graduate Student Researcher), Joy Bonaguro (Graduate Student Researcher), Paul Borokhov (Graduate Student Researcher), Stephen Carrier (Systems Administrator), Elie Daou (Visiting Graduate Student Researcher), Gunes Dervisoglu (Graduate Student Researcher), Magnus Fransson (Visiting Graduate Student Researcher), Timothy Hunter (Graduate Student Researcher), Raphaël Jatteau (Graduate Student Researcher), Matt Juhn (Applications Developer), Tasos Kouvelas (Postdoctoral Researcher), Tom Kuhn (R&D Engineer), Alex Kurzhanskiy (Research Engineer), Tristan Lall (Graduate Student Researcher), Pierre-Emmanuel Mazare (Visiting Graduate Student Researcher), Greg Merritt (Systems Analyst), Scott Myers (Applications Developer), Boris Prodhomme (R&D Engineer), Joao Rodrigues (Graduate Student Researcher), Samitha Samaranayake (Graduate Student Researcher), Mats Sandin (Visiting Graduate Student Researcher), Bill Sappington (Database Administrator), Krute Singa (Research Data Analyst), Nazy Sobhi (Business Analyst), Jerome Thai (Graduate Student Researcher), Luis Torres (Systems Administrator), Olli-Pekka Tossavainen (Postdoctoral Researcher), Joel VanderWerf (Systems Engineer), Fred Winik (Technical Writer), Dan Work (Graduate Student Researcher), Matt Wright (Graduate Student Researcher), Jeffrey Young (Communications Analyst), Yi Zhou (Undergraduate Research Apprentice).

Many people provided crucial subject matter during the project through conversations and interviews, original project documents, briefing materials, personal notes, emails, and other sources that made it possible to construct the narrative in this report. If, through our own oversight, we neglected to mention anyone who contributed to the project, we ask them to accept both our apologies and our gratitude.

EXECUTIVE SUMMARY

MOTIVATION

Traditionally, Caltrans' automated traffic data collection systems have relied on roadway-embedded sensors, such as loop detectors installed at fixed locations. With the recent prevalence of commercial traffic data sources, however, and the rising cost of maintaining state-sponsored traffic data collection operations, Caltrans is looking into purchasing probe data from the commercial sector. To help with this effort, PATH undertook the *Pilot Procurement of Third-Party Traffic Data* (Task Order 1) that investigated the feasibility and, under Task Order 2, the business case[1] for purchasing third-party probe data and fusing it with Caltrans' existing data for the purpose of estimating travel times. The intent was to demonstrate an efficient and cost-effective use of alternative traffic data sources to complement the detection systems currently installed and operated by Caltrans.

CONTEXT OF THIS STUDY

Traffic data is used to estimate current traffic conditions so that travelers and agencies can make better decisions about how to use and manage the transportation network. This contract explored the fusion of purchased probe data (vehicle speed and direction) with loop data (density, speed, and count) in the context of producing overall network speed and travel time estimates.

Speed and travel time estimates are useful in many circumstances, but current system control strategies (ramp metering, for example) require density data. The next phase of research will likely focus on new control strategy implementations that use fused probe and loop data for traffic management.

Our analysis of the ability to reduce the density of loops as presented in this report should therefore be viewed in the context of our current research, which focused on using fused probe data to estimate speed and travel time. The results and recommendations may differ when the requirement to control traffic is taken into account.

Additionally, future data procurement decisions need to determine whether the fusion and estimation steps would be done by an outside vendor or by transportation agencies.

OBJECTIVES

The objectives of this pilot project were to:

- Investigate the market for commercially available probe data
- Develop a process for obtaining probe data (unaggregated GPS point speeds)

---

[1] See the final report for Task Order 2, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods.*

- Purchase and integrate the data in a flow model for the purpose of estimating travel times
- Test the implementation
- Analyze the results

THE INVESTIGATION PROCESS

Our investigation proceeded through the following steps:

Assess current practices

A first step in investigating the purchase of commercially available probe data was to develop an understanding of the marketplace—its condition and current practices. We interviewed key players in the field of traffic data management from public agencies and industry and reviewed previous procurement efforts from both Caltrans and other state DOTs. What we found was a market undergoing rapid evolution and a widespread interest among stakeholders in integrating probe data with loop detector data. Stakeholders were seeking both accurate speed and travel time information (in addition to the volume and occupancy data currently available from loop detectors) and the ability to characterize corridor performance in more precise and comprehensive ways. This was seen as representing a valuable advance in the understanding and efficient operation of surface transportation systems.

Select sites

In order to purchase probe data, it was necessary to identify the highway segments for which data would be procured. We were particularly seeking locations where traffic was both variable and representative so that (i) the traffic state was not obvious or predictable, (ii) all ranges of speeds could be assessed, and (iii) the observations could be considered applicable to places with similar traffic phenomena. Using historical data from Caltrans, as well as historical and live data from Google Traffic, we found three segments with suitable traffic conditions: I-880 in the Bay Area, I-15 in Ontario, and I-15 in Victorville.

Establish ground truth

We also needed to get a picture of the actual traffic conditions at the selected sites (the “ground truth”) to be able to meaningfully assess the purchased data. We required independently observed travel times to use as a benchmark. To establish a reasonable estimate of ground truth, we used Bluetooth sensors which, when placed along the road, detect turned-on Bluetooth devices in passing vehicles. This gave us measured travel times of a sample of vehicles at each of the three sites, and it was these measurements that the probe data travel times would be compared to. At each location, we deployed ten sensors, spaced approximately one to two miles apart, for two weeks, thus gathering a total of six weeks of data at our three selected sites.

Solicit vendor proposals

A pivotal step in procuring traffic data from private vendors was the Request for Proposal (RFP) process: defining the data to be acquired, soliciting proposals from potential suppliers, reviewing responses, selecting a vendor, and contracting for product delivery. The purpose of the RFP was to explore the feasibility of purchasing unaggregated (unprocessed) data, the existing market for such data, its market price, coverage, and quality. Drawing on our interviews with stakeholders and analysis of previous procurement efforts, we developed and issued an RFP requesting unaggregated probe vehicle data, specifically point-speed data, for the three selected sites over a three-month period. We wanted unaggregated data so we could have better insight into the data's characteristics and use our own algorithms to generate travel times. Our RFP was testing the waters of the probe data market and, as such, required a balance of scientific rigor and simplicity—specific enough to get the data we wanted but not so complex that it discouraged vendors from responding.

Purchase data

The RFP succeeded in eliciting four vendor proposals, two of which substantially met the technical requirements, with acceptable costs. UC Berkeley's Purchasing Office wrote the contracts based on bids received from the vendors, and each side's legal department reviewed the contract. After discussions on data use, confidentiality, and price, the contracts were executed and the data procured.

Filter and assimilate the data

The probe data from each vendor was passed through a series of filtering steps which verified that the required data fields were present, removed duplicates and outliers, and mapped the point-speed locations to the road network. The data was then fed into the Mobile Millennium highway model which generated velocity maps and travel times. These could then be compared to the travel times collected by the Bluetooth sensors to see how closely the probe data matched the estimated ground truth.

The Mobile Millennium system can accept data from traditional sources (such as occupancy and counts from loop detectors) and point-speed measurements from providers of probe data. This enabled us to evaluate the performance of the data sources both individually and when fused together.

Balance loop and probe data

By reducing the amount of data from any particular source that we use with the model, we were able to test different proportions of each and evaluate the results.

CONCLUSIONS

Our investigation led us to the following conclusions:

- **Data procurement**—While the market is continually evolving, unaggregated probe data is available and can be procured successfully through an RFP process. Although limited in scope and aimed at specific targets, the RFP we issued stimulated the market to respond and revealed that vendors had both the desire and capability to supply the requested data.

- **Speed data and travel times**—Commercially available GPS point-speed data is usable for the intended application (speed data and travel times over a highway network) and can be successfully processed with the Mobile Millennium system to map velocities.

- **Data quality**—The quality of purchased probe data can be measured and compared to ground truth. We also found that when using probe data with a flow model on a highway, it is better to have less frequent data from a larger number of unique vehicles (a high penetration rate) than to have more frequent data from a smaller number of unique vehicles (a high sample rate). There could be other applications, such as arterial estimation, where high-frequency data would be better.

- **Data fusion**—Probe data can be successfully fused with loop detector data, and meaningful comparisons can be assessed.

- **Traffic estimation**—The accuracy of traffic estimation can be improved by combining probe data with loop detector data; even sparse probe data is useful. When probe data and loop data are fused together, average travel times can be estimated at an accuracy within the bounds of driver variability. For the purpose of estimating travel times, probe data can be used as a substitute for data from loop detectors.

- **Confidence in the model**—The quantitative and graphical results of the research give us confidence in both the modeling approach to roadway estimation and the effectiveness of the Mobile Millennium highway model for assessing and fusing procured data.

IMPLICATIONS AND FUTURE DIRECTIONS

Procuring commercially available probe data is just one of many important steps toward improving mobility in California by leveraging new data sources. This pilot project has shown that third-party probe data can be purchased commercially, fused with Caltrans' loop detector data, and used to improve traffic estimation along California's roadways without building new detector infrastructure. Probe data could even potentially fill in gaps where no loop detectors exist and speed data alone is sufficient.

Several implications follow from these findings:

Reduced dependency on loop data

While current system control strategies, such as ramp metering, require density data, it seems difficult to significantly increase and maintain the quantity of loop detectors on California roads. At the same time, the penetration rate of probe data is continually increasing and far from reaching its limits. This represents a sea change in the types of data available for traffic management and offers the prospect of migrating away from exclusive dependency on loop detectors over time.

Outsourced data collection

Purchasing probe data from the commercial sector means, in effect, outsourcing the collection of traffic data. Any such undertaking comes with new risks (e.g., data quality, privacy protection, business continuity) which would need to be managed through, for example, a careful vetting and data acquisition process and data assessment tools, processes, and standards.

Redesigned information systems

The research work in this task order was predicated on having a data assimilation and state estimation system in place that would allow the implementation, testing, and analysis of data hybridization. This required:

- Building and calibrating a model to estimate speed and travel times from probe data
- Developing a set of methods to fuse probe and loop detector data
- Creating tools to visualize the data
- Testing the tools and methods on real data from pilot sites
- Building tools to determine the quality of the methods, models, and data

Creating this mathematical and technology infrastructure points the way to the redesign of information systems that would make it possible to implement data fusion and take full advantage of hybrid data in traffic management systems.

New detector strategy

While further research on the position and spacing of loop detectors is needed, our initial results suggest that using the most critical detectors (those that add the most information value) rather than enforcing a minimal spacing between detectors could help optimize the existing stock of detectors and allow Caltrans to selectively focus maintenance efforts or supplement loop data with probe data when certain loops fail.

Augmented traffic measurements

This project studied the use of probe data for estimating travel times. However, the enhanced modeling and estimation accuracy demonstrated by data fusion also lays the foundation for better control strategies and operational decision-making. Augmenting traffic volume measurement with probe data, for example, could be a fruitful area for research in the future. Fused loop and probe data ("hybrid" data) could thus provide a pathway to the development and use of additional traffic measurements, such as arterial estimation, origin-destination information, demand modeling, and, ultimately, integrated corridor management.

Broader potential

Being able to reliably purchase and use accurate traffic data could potentially enhance many areas of interest to Caltrans beyond the efficient flow of goods and people across California, including transportation safety, work zone safety, emergency services, and evacuation management. The successful procurement and assimilation of third-party probe data represents an important step toward those possibilities.

Table of Contents

Chapter 1

Introduction

1 INTRODUCTION

This final report documents Task Order 1, *Pilot Procurement of Third-Party Traffic Data.* The task order falls under the *Hybrid Traffic Data Collection Roadmap* technical agreement number 51A0391, executed in September, 2009 under the parent agreement 22A0486 between the California Department of Transportation (Caltrans) and the Regents of the University of California (UC Regents).

This task order explored the fusion of purchased probe data (vehicle speed and direction) with loop data (density, speed, and count) in the context of producing overall network speed and travel time estimates. The pilot project was carried out by California PATH (Partners for Advanced Transportation Technology), a research unit of the Institute of Transportation Studies at the University of California, Berkeley.

1.1 MOTIVATION

Data is the lifeblood of effective transportation management. Traffic data is used to estimate current traffic conditions so that travelers and agencies can make better decisions about how to use and manage the transportation network.

Caltrans has traditionally captured traffic data from sensors buried in the road, such as loop detectors installed at fixed locations. While these detectors yield solid results for estimates of traffic volume and occupancy, they do not provide accurate travel time information unless the sensor coverage is very dense. In addition, roadway-embedded sensors currently in use are beginning to age, and the costs of maintaining or replacing them are high. Consequently, to avoid the capital and maintenance costs of new sensor technologies, Caltrans is looking into procuring good-quality data from third-party sources and integrating that data into Caltrans’ existing systems.

As part of that effort, PATH undertook the *Pilot Procurement of Third-Party Traffic Data* (Task Order 1) that investigated the feasibility and, under the related Task Order 2, the business case for purchasing third-party probe data (unaggregated GPS point speeds) from the commercial sector and fusing it with Caltrans’ existing data for the purpose of estimating travel times. The intent was to demonstrate an efficient and cost-effective use of alternative traffic data sources to complement the detection systems currently installed and operated by Caltrans.

1.2 NEW SOURCES OF DATA

From Roman roads to GPS-equipped probe vehicles, transportation engineers have built on the achievements of their predecessors. In 2008, the Mobile Century experiment demonstrated the feasibility of monitoring traffic by using data from GPS-enabled cell phones in a controlled environment. Beginning later that year, the Mobile Millennium project extended that capability to a complex urban environment on a far larger scale, gathering traffic information from the GPS in cell phones, processing it through a flow model, and distributing it back to the phones in real time to give users up-to-the-minute travel information. Since those groundbreaking achievements, the dramatic proliferation of smartphones and on-board GPS units carried by drivers has transformed ordinary cars, trucks, buses, and taxis into a mass of mobile traffic probes, a sea of data collection devices transmitting each vehicle's location, direction, and speed.

The massive availability of traffic data from GPS-equipped probe vehicles will have important consequences for both the traveling public and roadway operators, and Caltrans, in particular, is set to benefit tremendously from its use. Leveraging this data source could drastically cut the ongoing costs of traffic monitoring and expand coverage to thousands of miles of highways and urban arterials for which sensor installations are not considered an option. Moreover, it could be used for improving system management, traffic flow, transportation safety, work zone safety, emergency services, and evacuation management, as well as increasing the efficient movement of goods and people across California. Further, the dissemination of traffic information could promote a form of system "self-management" in which individual commuters can make informed travel decisions. Not only would each user benefit personally, but the entire driving community would enjoy more balanced loads across the road network.

1.3 A HYBRID TRAFFIC DATA SYSTEM

The data collected from cell phones and other GPS devices, however, is limited to velocity information. Caltrans has relied on fixed loop detectors to provide measures of traffic flow and occupancy rate, and current system control strategies (ramp metering, for example) in fact require density data. It is therefore unlikely that probe data will completely replace existing detection systems in the foreseeable future.

However, using mobile probe data to complement existing detectors—a so-called "hybrid" traffic data collection system—has real viability. In such a setup offering ubiquitous availability of speed information, loop detector stations would be needed only to maintain accurate flow counts. Larger spacing intervals could be allowed between stations, and therefore equipment could be deployed much more sparingly than it is today. Caltrans could thus be getting much more information while spending much less for it.

(It should be noted that this project focused on using fused probe and loop detector data to estimate *speed and travel time*, rather than to explore control strategy implementations. Using fused data for traffic control strategies will likely be addressed in future research.)

1.4 PROBLEM STATEMENT

Purchasing traffic data represents a fundamental shift for Caltrans that poses its own set of challenges. Therefore, in addition to the engineering work on integrating probe data with loop detector data, the research team on this project gave substantial attention to investigating the business aspects of traffic data procurement, including interviewing stakeholders in the transportation community, reviewing previous procurement efforts, assessing current practices, defining a procurement process, and so on. The task of acquiring third-party data and fusing it with Caltrans' existing data thus presented a number of key questions:

- Is probe data available in the marketplace?
- What data is available?
- How good is it?
- Can it be purchased? How? At what cost?
- Can it be tested/used/fused with loop detector data?
- How can this be achieved?
- With what results?
- To what extent can probe data supplement or supplant loop data?
- Can it be integrated into Caltrans' Transportation Management Systems (TMS)?

1.5 OBJECTIVES

To address these questions, the research team identified the following objectives for this pilot project:

- Investigate the market for commercially available probe data
- Develop a process for obtaining probe data (unaggregated GPS point speeds)
- Purchase the data
- Integrate the data in a flow model for the purpose of estimating travel times
- Test the implementation
- Analyze the results

1.6 CONTRACTUAL DELIVERABLES AND ORGANIZATION OF THIS REPORT

Task Order 1 categorizes the proposed work into eleven tasks:

- Task 1. Project management, coordination, and outreach
- Task 2. Assessment of current practices and lessons learned
- Task 3. Site selection and assessment
- Task 4. Development of technical specification
- Task 5. RFP solicitation and evaluation
- Task 6. Contract management and data procurement
- Task 7. Data collection rollout
- Task 8. Test fusion schemas
- Task 9. Data fusion
- Task 10. Research on hybridization of multiple data sources in Traffic Management Systems
- Task 11. Final report

This report is organized into seven chapters and addresses the Task Order work breakdown as follows:

- Chapter 2, *Assessment of Current Practices*, addresses the work of Task 2. It surveys current practices among stakeholders in the traffic management community, both public agencies and private industry, focusing particularly on their need for data and its intended uses. The chapter also includes analysis of previous data procurement efforts by Caltrans and other state DOTs.

- Chapter 3, *Choosing a Site and Establishing Ground Truth*, details work done for Task 3. It describes the process of selecting the sites for which to acquire data and the elements considered in making a choice. It also describes the effort to develop a reasonable estimate of the actual traffic conditions (the "ground truth") at the selected sites using Bluetooth sensors, in order to obtain a benchmark for judging the procured data.

- Chapter 4, *Procuring Traffic Data*, speaks to Tasks 4, 5, and 6. It details the efforts to procure data from commercial vendors, in particular the issues, considerations, and decisions involved in developing a Request for Proposal and evaluating the responses. Included are recommendations to Caltrans based both on that experience and on the research described in Chapter 2.

- The work of integrating, fusing, and evaluating data from multiple sources is presented in Chapter 5, *Data Fusion*, which addresses the work of Tasks 7, 8, and 9. The chapter describes the mathematical context of data assimilation, the metrics used to quantify the usefulness of probe data, comparisons between different data sources, and the results of combining various amounts of probe and loop detector data.

- Research into enhancing Traffic Management Systems to use multiple data sources (Task 10) was subcontracted as part of this task order. Chapter 6, *Hybridization in Traffic Management Systems*, describes the subcontracting process and includes the subcontractor's final report.

- This final report (Task 11) concludes with Chapter 7, which reviews the project, summarizes the results and the conclusions that emerged from the work, and considers the implications for the future.

In addition, the numerous management activities needed to guide the project to completion fall under Task 1. We are grateful to Caltrans for the close collaboration, ongoing consultation, and direct input—in the form of meetings with Headquarters and District staff, group presentations, weekly conference calls, and other communications—that they provided in the course of this work.

Chapter 2

Assessment of Current Practices

A first step in investigating the purchase of commercially available probe data was to develop an understanding of the marketplace and the current practices in the traffic management community. What we found was a market undergoing rapid evolution and a widespread interest among stakeholders in integrating probe data with loop detector data. This chapter describes the investigative process we undertook and what we learned from the experience.

1 ASSESSING CURRENT PRACTICES

With the recent prevalence of commercial traffic data sources and the rising cost of maintaining state-sponsored traffic data collection operations, many state transportation agencies including Caltrans are looking into purchasing probe data from the commercial sector. To aid in this endeavor, PATH pursued two lines of inquiry to identify current practices:

- **Stakeholder outreach**—We conducted stakeholder outreach to learn the impact Caltrans' purchasing probe data would have on stakeholders' day-to-day operation, practice, policy, and management. Specifically, we aimed to gather information about the following issues:
    - Current practice of traffic applications
    - Current practice of existing and candidate traffic data sources
    - Current practice of traffic information dissemination
    - Preferences for traffic data sources
    - Recommendations and concerns about the quality, accuracy, and availability of traffic data
    - Traffic system management operations and challenges
    - Visions and strategies for traffic system management
- **Analysis of previous RFPs**—We reviewed previous procurement efforts, both from Caltrans and from other state DOTs, to see what could be learned from that experience.

2 ABBREVIATIONS USED IN THIS CHAPTER

ADT- average daily traffic

ATMS - advanced traffic management systems

BTS - Berkeley Transportation Systems

CAD – computer-aided dispatch

CCTA - Contra Costa Transportation Authority

CMA - Congestion Management Agency

CMP - Congestion Management Program

CMS - changeable message signs

DOT - Department of Transportation

GPS - global positioning system

HOV - high-occupancy vehicle

HQ - headquarters

IFB - invitation for bids

ITS - intelligent transportation systems

MPH - miles per hour

MPO - metropolitan planning organization

MTC - Metropolitan Transportation Commission

OD - origin–destination pairs

PATH - Partners for Advanced Transportation Technologies

PeMS - Performance Measurement System

QA - quality assurance

RFI - request for information

RFP - request for proposals

SANDAG - San Diego Association of Governments

TIMC - Transportation Incident Management Center

TMC - Traffic Management Center

TMDD - Traffic Management Data Dictionary

TOPL – Tools for Operational Planning

TTI - Texas Transportation Institute

3 STAKEHOLDER OUTREACH: RESOURCES AND METHODS

PATH identified key players from industry, public agencies, vendors, and other stakeholders in the field of traffic data management. Due to time limitations, we selected only a few members from industry and public agencies for interviews. Those interviews were conducted mostly in person, with some conducted over the phone. The questions were open-ended to maximize the input of the interviewees.

3.1 GOVERNMENT ORGANIZATIONS

We interviewed the following public agencies:

- **Caltrans**—We interviewed key Caltrans personnel at several locations:
    - Caltrans Headquarters, Division of Research and Innovation
    - Caltrans Headquarters, Division of Operations
    - Caltrans District 4
    - Caltrans District 7

  We chose members of Caltrans Headquarters to interview because of the organizations' position as decision makers. We also interviewed members of the office of operations in the two largest Caltrans districts (District 4 in Oakland and District 7 in Los Angeles) to identify the impact purchasing probe data would have on their day-to-day operations.

- **City of San Jose Transportation Department**—We interviewed members of the City of San Jose Transportation Department to gain insight into a large city with experience managing a Traffic Control Center. The department is responsible for services that provide for the safe and efficient movement of people and goods. The department plans and programs capital improvements for vehicles and pedestrians, and their engineers and planners work closely with other agencies to ensure that transit services and freeway improvements meet the needs of San Jose residents and businesses.

- **Contra Costa County Transportation Authority (CCTA)**—This organization, which serves as the Congestion Management Agency (CMA) for Contra Costa County, has experience in purchasing probe data. As the CMA, the agency must, under State law, prepare a Congestion Management Program (CMP) and update it every two years. The CMP is meant to outline the CMA's strategies for managing the performance of regional transportation operations within the county.

- **Metropolitan Transportation Commission (MTC)**—MTC is the transportation planning, coordinating, and financing agency for the nine-county San Francisco Bay Area. Over the years, state and federal laws have given MTC an increasingly important role in financing Bay Area transportation improvements. MTC functions as both the regional transportation planning agency (a state designation) and, for federal purposes, the region's metropolitan planning organization (MPO).

3.2 PRIVATE SECTOR (SYSTEM INTEGRATORS)

We interviewed the following private firms:

- **Berkeley Transportation Systems** (a division of Iteris Inc.)—Berkeley Transportation Systems (BTS) is the original designer and developer of PeMS. They provide software and transportation data experience, as well as operating and maintaining the Caltrans version of PeMS. BTS currently has seven different PeMS deployments around the world collecting and analyzing traffic data, each with a variety of algorithms.
- **Delcan Corporation**—Delcan offers a variety of Intelligent Transportation Systems (ITS) services covering all phases of the project cycle, from requirement workshops and project design, through evaluation and procurement of all hardware and software. This includes software development, configuration management, testing, integration, training and documentation, troubleshooting, and maintenance support.

3.3 EXAMPLES OF INTERVIEW QUESTIONS

- What is the demand for traffic data from different applications?
- What kinds of applications can function based on probe data only?
- What kinds of applications cannot function on the probe data only and need additional types of data such as loop data?
- Is additional probe data useful for the organization?
- Are there any preferences for the data sources or type of data?
- How and where can probe data complement the loop data?
- What strategies exist to enhance existing traffic system management approaches?
- What is your vision and strategy?

3.4 SUMMARY OF RESULTS

Historically, Caltrans has relied on a traffic data collection network constructed largely of roadway-embedded sensors, i.e., loop detectors installed at fixed locations. While this method yields reasonable results for estimating volume and occupancy, it does not provide accurate travel time information unless the sensor density is significantly increased.

- Probe data could provide a valuable source of data for accurate travel time information. However, volume and occupancy are necessary for the following applications in Caltrans operations:
  - HOV lanes
  - ramp metering
  - incident management
  - congestion management
  - construction and lane closure
  - census program
- In general, stakeholders believe that traffic management via the integration and fusion of various data sources represents an evolution in the understanding and efficient operation of surface transportation systems. It is no longer possible to solve traffic congestion by simply adding more lanes to roadways. A new age of traffic management is emerging that will shape operations in the decades to come.
- Purchasing probe data could provide Caltrans traffic operations decision makers with a unique opportunity to acquire knowledge about the rapidly evolving field of traffic data management as well as to set the strategic direction of policies on the future use of traffic data.
- Probe data could present a valuable source of data in rural areas where loop data is absent. Also, probe data could complement loop data when the loops are not reporting.
- All agencies appreciate more and reliable data regardless of data source. The abandonment of loop data will only affect the Traffic Management Center (TMC) systems that are a function of speed and ramp metering.
- Ideally, system integrators would like to see a combination of data sources that effectively close data gaps (such as video detectors, radar detectors, and probe data) but acknowledge the importance of lane-by-lane data for incident management and measuring travel time differences between HOV lanes and mainline lanes. Furthermore, lane-by-lane sensors are important when doing detailed lane-by-lane systems or ramp metering. Ultimately, Advanced Traffic Management Systems (ATMS) in California could be enhanced by increasing probe data, because the lack of reliable detection is currently hindering systems (missing/malfunctioning detectors make it difficult to measure travel times and congestion).

- Volume and speed data are both needed to characterize corridor performance and indicate the amount of travel and congestion. It is also necessary to identify where break points in speed are (points at which speed starts to decrease from free-flow) in order to manage bottlenecks. If purchased probe data can identify these issues, then it would be useful. The hybridization of data would also work if compromises made in the data fusion don't compromise the precision of the data.

4 MEETINGS WITH STAKEHOLDERS

This section summarizes our interviews with key members of various organizations. For a summary of the interview responses grouped by category (Loop data, Data Functions, Data Sources, Third-Party Provider, and Solutions), see *Grouped interview results* on page 23.

4.1 BERKELEY TRANSPORTATION SYSTEMS (A BUSINESS UNIT OF ITERIS)

Many PeMS applications can be used without loop data, particularly those that are most frequently used. These applications use data measures like travel time, trends in travel time, and facility degradation. Bottlenecks are also a function of speed and therefore do not require loop or flow data. There are, however, many PeMS applications that require flow data, including measurements of delay, congestion, the size of queues from incidents or lane closures, and the cost of delay. This data is not solely reliant on loops but can also be obtained by using different types of raw data, including Fastrak data, which provides samplings of paths and people.

TMC software is capable of controlling messages on changeable message signs (CMS) and includes all the steps needed to manage an incident on the freeway (calling tow trucks, police, coordination with hazmat teams, traffic diversion). However, the majority of TMC software is useless because most of what traffic management centers do does not require loop data or any data at all. While information on traffic maps is derived from loop data or other point system data, the map is not used to manage anything. Similarly, TMCs are not responsible for changing ramp metering, and thus their software does not require loop data. For example, a city in District 7 requested that the ramp meters be turned off during a heavy shopping period to decrease congestion on local streets, but congestion then increased on freeways. TMC controllers were not responsible for figuring out the cause of congestion; it was engineers who used loop data to determine the cause of congestion. Engineers use the information from their data studies to instruct TMCs how to operate, while TMCs rely on people-to-people communication.

Caltrans District 10 uses a software system called IRIS (part of TMC) which controls cameras and issues automated warnings (for example, if fog is worse than a given value, the warning is issued). This type of software is more suitable for the duties of TMCs. Conversely, Caltrans District 4 recently had speed information sensors installed by MTC, and this data is given to Caltrans but it is unknown whether this data is put to use, affirming one interviewee’s response that the collection of data and its application in traffic systems are not directly correlated. As such, the abandonment of loop data will only affect the TMC systems that are a function of speed and ramp metering systems, but currently the data that is derived from these systems is not put to use by TMCs (it may be used by engineers as seen in District 4).

In response to the question regarding the importance of Caltrans knowing how data is fused, one interviewee suggests from his own experience that it may be irrelevant. The interviewee has described the functions that fuse data in PeMS applications, but Caltrans is not convinced that PeMS is doing the right thing. The interviewee does note the validity of Caltrans' concern over the accuracy of data projections.

4.2 DELCAN

Delcan currently uses probe data extensively outside California in states such as Michigan and New Jersey. The company gets travel times for CMS systems from third parties like NAVTEQ or Traffic.com. Delcan is used to having to deal with little data and has procedures to determine travel time from probe data and even cellular probe data, although it is more desirable to have GPS probe data for greater accuracy. Ideally, Delcan would like to see a combination of data sources that effectively close data gaps like video detectors, radar detectors, and probe data but acknowledges the importance of lane-by-lane data for incident management and for measuring travel time differences between HOV lanes and mainline lanes. Furthermore, lane-by-lane sensors are important when doing detailed lane-by-lane systems or ramp metering. Ultimately, ATMS systems in California could be enhanced by increased probe data because the lack of reliable detection is currently hindering systems (missing/malfunctioning detectors make it difficult to measure travel times and congestion). Delcan can improve/expand their ability to post travel times by increasing probe data (this is also a low-cost solution). Currently, Delcan has its own algorithm for calculating travel time and congestion levels from Speedinfo stations that collect speed data.

4.3 CALTRANS HEADQUARTERS, DIVISION OF RESEARCH AND INNOVATION

We interviewed two key members in this office: interviewee A and interviewee B. Interviewee A wants information from third-party data providers, not unaggregated data, because they want information that's needed to make decisions about system management. This falls in line with what interviewee A identifies as the main objectives of Caltrans (system management and meeting traveler needs). If information is provided to travelers, it is assumed that they will make better decisions that will help manage the system without Caltrans having to take action. Interviewee B expands upon interviewee A's identification of Caltrans' main objectives, claiming that Caltrans' main mission is to provide travelers with a safe and efficient trip by minimizing the aggregate cost of travel time and congestion. As such, travel time is the ultimate objective of Caltrans according to interviewee B.

The main issue that arises with purchasing information compared to unaggregated data is the validity and accuracy of information provided. If information is purchased, both interviewees agree that some sort of insurance is needed to address the validity, accuracy, reliability, and timeliness of this information. Interviewee B addresses concerns among local districts regarding PeMS applications because they don't know how data is fused, and there have been instances where data has proven to be invalid. Consequently, interviewee B believes there is a growing need to find a way of doing quality assurance (QA) to insure that information is actionable and thinks that Caltrans should have the ability to check data that received information is founded on.

Interviewee B identifies a pool-funded project among multiple states and actors (Texas, TTI, Alaska, Pennsylvania, Michigan, Virginia, and the federal government) as an example of checking data and insuring it has adequate quality assurance. According to this project, there are two technologies available to do quality assurance. The first is to run probe vehicles, which interviewee B identifies as the more expensive option ($20/data point). The second is through Bluetooth readers (0.2 cents/data point). Baseline data, collected either by states or a hired contractor, is another option, but interviewee B identifies a problem with this method. It would be difficult to insure there is no collusion between various contractors in verifying data and information. An alternative would be to hire an academic contractor because they are more independent, but interviewee B suggests that they do not always have a valid perspective on practical things (for example, spending a huge amount of time and money to collect a very small amount of data).

One of the main arguments for keeping loops is the data requirement of ramp metering. Interviewee B argues that ramp metering does not necessarily need flow and occupancy on the mainline, but you do need to know if a vehicle is queued at the ramp and if the ramp is full so traffic doesn't back up into the city. Some type of presence detector will always be needed locally at the intersection and at the ramp meter, but it does not necessarily have to be in the form of loops. Alternatives include camera systems, Sensys nodes, and radars. Interviewees also acknowledge that many ramp metering rates do not change based on loop data information but rather on the time of day. Another main argument in favor of loop data is the need to measure flow in HOV lanes compared to main flow lanes. Differentiating between the two is important because of speed differences that affect travel time. Interviewee B suggests that most GPS systems today can achieve 1–2 meter accuracy, which would account for lane accuracy. INRIX and other data providers could provide this service with up to 70–80% accuracy.

Interviewee B also extensively discussed methods of figuring out travel time. A more valid way of doing travel time, he suggests, is by setting way points (NAVTEQ uses this in the form of virtual trip lines similar to the technology used in Mobile Millennium). This meets Caltrans' objective of providing information to allow travelers to make better decisions regarding travel. Interviewee B also notes that the most useful place to display travel times is at link boundaries, so travelers can make the decision to continue or get off. Another issue concerning travel time is the ability to provide the public with future travel times. This can be done heuristically or through an analytical model, but both methods need Origin–Destination (OD) pairs, and no one has it. Origin–Destination pair is the key variable in predicting future travel time and can be derived from Bluetooth technologies that link up with a Mac address. The advantages of using Bluetooth technologies include increased travel time accuracy, OD pair availability, and reduced expense.

Interviewees identify one of the main reasons Caltrans has decided to research the procurement of data as an increase in loop degradation. Interviewee B argues that it is not the loops that are not working but rather the middleware (modem card, front-end processor, ATMS, the 170 controller) that is outdated and hinders loop data collection and efficiency. Interviewee B provides a solution to the middleware problem by examining District 10's loop efficiency. Two years ago, 97% of loops were active, whereas six months ago only 67% were active due to the removal of specific middleware.

4.4 CALTRANS HEADQUARTERS, DIVISION OF OPERATIONS

Background: Caltrans wanted to start purchasing third-party data years ago and received a couple million dollars from the legislature to do this a few years ago. Funding was obtained because of a promise to the legislature that the funds would go towards replacing in-ground detection suffering from maintenance and safety issues, but none of the proposals met the requirements of the RFP so the project was canceled. (For further details, see *Analysis of previous Requests for Proposals* on page 29.)

Current situation: When in-ground detection first started, it was estimated that loops would be installed every half mile. Caltrans would now like to place loops only where they are necessary (ramps, for example). Rather than replacing broken loops, Caltrans would like to augment or extend the loops they already have without installing additional in-ground detection. Speed data is the least valuable information that comes from in-ground detection, according to the interviewee. This sort of information is more relevant to operations, but in the planning realm, Average Daily Traffic volumes (ADTs), volume, and occupancy have a higher value. Occupancy is needed because the Federal government pays for the operation of HOV lanes, and they must perform in a certain manner or funding will be revoked. Occupancy is also needed in the mixed flow lanes because if Caltrans goes to active traffic management, speeds change based on lane performance.

With regard to PeMS, BTS maintains the system/software and Caltrans maintains the hardware (it was necessary to have the hardware within Caltrans firewalls). Caltrans is satisfied with PeMS because data is consistent and you can do things with it (compared to data from third parties). The creation of PeMS has also helped increase communication and has established a data repository with the historical data the interviewee seeks.

The interviewee thinks the initial purchases from third-party data collectors will be speeds and travel times to use as fill data. If speed data cannot be purchased, then the project is over. Caltrans does not want to compete with the private sector for 511 information because the private sector does it better. The interviewee thinks that providing travel time is not the main role of Caltrans, but good data is needed for planning purposes (in relation to ADTs and delay) and operations (real-time data, TOPL Simulation tool). As such, the interviewee is not impressed by the recent I-95 project because it only uses travel times purchased from INRIX. This information is not hybridized with any other data.

The interviewee would prefer unaggregated data because the data can then be hybridized and put into PeMS. Third-party data collectors often use algorithms that smooth data, resulting in inaccuracies when conditions change (rain, incidents). The interviewee is willing to buy aggregated data if it is known where the data is coming from and how it is aggregated. The main issue is being able to verify its validity and reuse the data. For example, you can buy data from INRIX one year, but that data will not be available the next in an archived format. The interviewee sees historical data as necessary for planning purposes.

Future: At some point, Caltrans would like to be able to calculate volume from various data collection mechanisms as they penetrate the market. Caltrans would also like to work more extensively on multi-modalism. This is already apparent in District 11 (San Diego) because SANDAG is responsible for the train and bus system in addition to managing the freeways.

If this project yields multiple proposals from different vendors, then the usefulness of each will be determined and purchased where appropriate. The interviewee doesn't foresee one company covering the whole state, and there are no specific corridors in mind (I-15 and I-80 were used as examples). If responses to the RFP are lackluster, Caltrans will look more closely at Bluetooth technology. Bluetooth technology was being explored further in District 5 (Santa Barbara to Santa Cruz) through Task Order 7 that would allow for the installation of Bluetooth readers to provide the district with speed and travel times for CMS. (Unfortunately, Task Order 7 was cancelled due to contracting issues.)

The interviewee also noted (as did San Jose Transportation Department) that it is necessary to increase communication. It is an important aspect of both planning and operations that people get along and include others in decision making.

4.5 CALTRANS DISTRICT 4

Data representing flow and volume counts are necessary to manage road construction and lane closures. Similarly, lane-by-lane data is needed to measure the HOV lane. Caltrans District 4 has a specific need for truck classification information, an increase in accuracy for ramp metering and congestion monitoring, data measuring lane volume and speed, data for occupancy (algorithm needs to be written), and data to measure lane quality. The District finds probe data suitable when there is no loop data available but only for traveler information and changeable message signs.

4.6 CALTRANS DISTRICT 7

Currently, District 7 does not use probe data, and most traffic management applications rely on detection data from loop detectors that measure occupancy and volume and allow for speed calculation. Incident detection, congestion management, and ramp metering use occupancy and flow data from loop detectors. Occupancy and volume play a large role in determining the parameters for algorithms which help determine performance and saturation of the freeway.

District 7 uses travel times for changeable message signs, and travel times are published on 511.org. Travel times are mostly determined from loop detection data and also, to a lesser extent, from Speedinfo. District 7 is more flexible when it comes to alternative means of determining travel times because the technologies are widely available and mostly reliable. Speed information is imported into ATMS through Ethernet to calculate travel times.

Current incident detection measures use real-time data, not historical data. The data requirement is thus higher. The combination of probe data and historical data could work to allow for an educated guess at incident and congestion detection, but District 7 is not currently using historical data. It is possible that probe data could play some role if real-time calculations can be built in conjunction with historical data, but more computing power would be needed to store all this data. Probe data could also be used as a data trigger to validate scenarios without volume/occupancy data.

In rural areas ramp metering is not centrally controlled, and locally controlled metering is based on experience of peak times. District 7 cannot currently run detection algorithms for incidents or congestion in rural areas because there aren't enough loops.

District 7 Traffic Management Centers currently use a Delcan system from which District 4 also receives data. The Delcan system receives its data from loop detectors. The interviewee believes that the algorithms used are adaptable to include probes, but it has not been done. The software is capable of receiving external data and applying algorithms to that data, but the current Speedinfo data cannot be used for incident detection or congestion management because it does not provide information for these detections.

District 7 currently has nearly 10,000 loops in the mainline (15,000 loops in total including those on ramps) and is trying to find a substitute technology for loops because they're costly to maintain. The district has looked into radar, Sensys nodes, and other off-the-shelf technologies, but these have their shortcomings (radar strength deteriorates, Sensys stations have battery problems).

4.7 SAN JOSE DEPARTMENT OF TRANSPORTATION

The county currently collects freeway and arterial performance data as well as major intersection information, which is shared across multiple agencies. A database of all projects that collect data is maintained and archived and is used for development decisions, signal timing, and signal modifications. There is a compilation of data coming through many projects and regional programs. Often, nothing is done with the real-time data that is collected.

At the time of this interview, the San Jose Transportation Department was working on two major projects:

1. **Increase infrastructure**, using $20 million in funds obtained through Prop 1B:

    - Traffic light synchronization program
    - Replacing/upgrading 900 controllers
    - Installing over 50 miles of fiber backbone and a wireless communications system that allows for the transmission of video and traffic data through Ethernet
    - Adding 150 cameras citywide to the current 100 cameras at major arterials where there are congestion complaints
    - Adding adaptive traffic controllers to address seasonal traffic around the airport, downtown, and mall areas
    - Predictive priority for tracking bus arrivals (already implemented); no legitimate problems with fire or police departments thus far

    All these updates were to be complete by the end of 2012.

2. **TIMC (Transportation Incident Management Center),** in the final stages of conceptual operations and expected to be complete by late 2013. Its purpose is to monitor and manage traffic conditions while capturing and using data in a more efficient manner. Part of this project also includes the expansion of the communications network to more efficiently share video across the county. Data is received through fiber-optics and copper. TIMC will take San Jose from a reactive to a more proactive role, but many things still stand in the way of maximizing efficiency, including a lack of resources.

   TIMC also plans to develop an Events Tracking System, which would integrate different data systems and people. Siemens has been hired to go through the systems engineering process. It is unknown, so far, who will fuse all the data that results from this system. San Jose doesn't think it is realistic for data from Caltrans or outside organizations to be easily integrated into this type of system, introducing the importance of information flow through people- to-people communications.

   There is more responsiveness in person-to-person communication, and San Jose recognizes the importance of human factors in the collection and relay of data. The city is interested in sharing information between systems, specifically to the CAD system (fire and police dispatchers). They are also interested in implementing procedures/provisions that permit the easy transmission of information between contractors/inspectors and operations staff (the main people interacting with the system).

   There is also a safety component to the TIMC that will evaluate increased efficiency and safety through a reduction in crashes/incidents. This will eventually lead to multiple RFPs that will be handled by Siemens.

There has not been a significant focus on *live* data monitoring, reviewing, and analyzing in either project. It has not been made a priority. San Jose is currently heavily reliant on camera systems and loop data for assessing traffic congestion, but the above two projects aim to integrate the use of data into the process.

INRIX provides congestion mapping, arterial congestion mapping using smart-ware technology, ability to historically compare travel times and congestion, and collection of GPS and probe data. San Jose is looking to integrate this type of information into their traffic systems. Volume information, however, is not included, so the city is looking into supplementary data sources (embedded loops, vehicle detection technologies, Bluetooth) to include this data.

4.8 CONTRA COSTA TRANSPORTATION AUTHORITY

The Contra Costa Transportation Authority (CCTA) is a public/private partnership that employs 19 staff members. Consultants, not new staff members, are hired when new projects surface, which allows for greater efficiency and a decrease in spending. Consultants typically charge $150/intersection for each peak period ($300 for one day of intersection counts). CCTA receives funding primarily from the Federal government and from Measure J. The agency is required to collect transportation data for two programs, the county-wide plan and the travel demand forecasting model. This data is collected every 5–10 years manually by consultants, Caltrans floating car runs, and through PeMS. Recently, CCTA attempted to use loop data for floating car runs using a program written by BTS. A robust set of data resulted, but because of inoperable detectors CCTA will probably return to normal floating car runs. Data is also required for the state-mandated congestion management program that requires the observance of the CMP network of freeways and principal arterials (this program looks at the level of service and compares it to service standards set for the network). All data is kept in house on a server.

CCTA recommends that if data is purchased, it should come directly from the source and not the server (where it can be manipulated). Furthermore, CCTA believes that Caltrans should continue to provide data to third parties at no cost because it is consistent with Caltrans' goal of enabling travelers to be both informed and efficient. Data sharing also increases transparency. CCTA identifies loop detectors as the standard in data collection (particularly for planning) but also acknowledges their high cost and maintenance issues. CCTA also indicates that many loop detectors are not in the right locations for future ramp metering and suggests they should be moved further out and properly coordinated with the mainline. Alternatives to loops, if technology allows for it, could include vehicle integration and connected car technologies.

4.9 METROPOLITAN TRANSPORTATION COMMISSION (MTC)

We interviewed two key members in this organization: interviewee A and interviewee B. Through the 511 Task Order, interviewee A is evaluating options for procuring new data. This involves an evaluation of what the data requirements are and what vendors are available. Interviewee A is not sure if the data procurement process would meet everyone's needs, as 511 primarily needs speed data. The Arterial Operation group and the planning department need data from in-ground detection that measures volume and occupancy. They are most likely going to be looking at probe data and spot speed data to generate travel times but will not be procuring travel times, only speed data. Currently, 511 receives free data from Caltrans. They generate their own data through vehicle re-identification, and they purchase data from SpeedInfo ($110/month per unit). In July 2012, they released an RFI to solicit information from vendors based on requirements for 511 and the Arterial Operation group and depending on the results, an RFP will be conducted. The next step would be to sunset the vehicle re-identification system and the Speedinfo contract. A parallel effort would also be made to research Bluetooth readers in call boxes (feasibility is still unknown depending on battery size, solar panel size, and so on). While call boxes have 20% of their call volume from 10 years ago, about 20,000 calls are still made each year.

Interviewee B is currently working on a wide range of system management items (with Caltrans District 4 on ATMS) that need large amounts of traffic data. Interviewee B indicates that his needs are in line with the needs of ATMS. Volume and speed data are both needed to characterize corridor performance and indicate the amount of travel and congestion. Data conveying where break points in speed are is also necessary to manage bottlenecks. If data is purchased that can address these issues, then it is useful by interviewee B's standards. Interviewee B thinks the hybridization of data would also work if the process of data fusion don't compromise the precision of the data.

5 GROUPED INTERVIEW RESULTS

The following sections group the interview responses into:

- Loop data
- Data functions
- Data sources
- Third-party data providers
- Solutions

5.1 LOOP DATA

5.1.1 BTS

- Many of PeMS applications can be used without loop data, particularly those that depend on travel time.
- The TMC software has much more capacity than it is used for, including loop data uses which are resolved elsewhere.

5.1.2 CALTRANS HQ

- Loops are currently needed for ramp metering.
- HOV lane management currently requires loop data.
- In one district, 1/3 of loop detectors became dysfunctional in an 18-month period due to middleware problems and not the loops themselves.
- Loops should only be added where necessary and not to the 0.5 mi density originally envisioned.
- Speed data is the least valuable data that comes from loop detectors.

5.1.3 CALTRANS DISTRICT 7

- Most traffic applications rely on loop data.
- Loops provide data for travel times for CMS.
- District 7 currently has nearly 10,000 loops in the mainline (15,000 loops in total including those on ramps).
- District 7 is trying to find a substitute technology for loops because they’re costly to maintain.

5.2 DATA FUNCTIONS

5.2.1 BTS

- Some PeMS applications require flow data, including measurements of congestion and queue sizes.

5.2.2 CALTRANS HQ

- Travel time is the paramount objective.
- Some kind of presence detector will always be needed at the intersection and the ramp meter but not necessarily loops. Alternatives include cameras, Sensys nodes, and radar.
- In the planning realm, ADTs, occupancy, and volume data have higher priority than speed data.
- PeMS has created a data repository; historical data is important.
- Caltrans can't compete with the private sector for 511 data.
- Prefer unaggregated data that can be hybridized and put into PeMS.
- Would like to someday collect volume and work toward multi-modalism with external data.

5.2.3 CALTRANS DISTRICT 7

- Most traffic applications rely on loop data for occupancy, volume, and speed.
- Incident detection, congestion management, and ramp metering use occupancy and flow data from loop detectors.
- District 7 cannot currently run detection algorithms for incidents or congestion in rural areas because there aren’t enough loops.

5.2.4 MTC

- Volume and speed data are needed to characterize corridor performance and indicate the amount of travel and congestion.
- Data identifying where break points in speed are is also necessary to manage bottlenecks.

5.2.5 CCTA

- Data required for two programs: the county-wide plan and the travel demand forecasting model.

5.2.6 DELCAN

- Use data for CMS systems.
- Lane-by-lane data needed for incident management and for measuring travel time differences between HOV lanes and mainline lanes as well as for ramp metering purposes.
- Has developed its own algorithm for calculating travel time and congestion levels.

5.3 DATA SOURCES

5.3.1 CALTRANS HQ

- Wants processed data from third parties, not aggregated data.
- Presence detector will always be needed at the intersection and the ramp meter but not necessarily loops. Alternatives include cameras, Sensys nodes, and radar.
- Waypoints (similar to what NAVTEQ uses in the form of virtual trip lines and which were used in the Mobile Millennium project) are best way to provide travel time data.
- Bluetooth technology is the best way to provide future travel times.

5.3.2 CALTRANS DISTRICT 7

- ATMS and Ethernet used for calculation of travel times.
- District 7 Traffic Management Centers currently use ATMS (Delcan system) from which District 4 also receives data.
- The ATMS receives its data from loop detectors.
- The ATMS is capable of receiving external data and applying algorithms to that data, but the current SpeedInfo data cannot be used for incident detection or congestion management because it does not provide information for these events.

5.3.3 SAN JOSE DEPARTMENT OF TRANSPORTATION

- INRIX provides congestion mapping, arterial congestion mapping using smart-ware technology, ability to historically compare travel times and congestion, and collection of GPS and probe data.
- San Jose is looking into supplementary data sources (subsistent loops, vehicle detection scenarios/technologies, Bluetooth).

5.3.4 CCTA

- Data should come directly from the source and not the server where it can be manipulated.
- All data is kept on in-house server.
- Loops detectors are the standard in data collection for planning.
- Floating car runs.

5.3.5 DELCAN

- SpeedInfo Station speed data
- Probe data
- Travel times from NAVTEQ or traffic.com

5.4 THIRD-PARTY DATA PROVIDERS

5.4.1 BTS

- Third-party data collectors do not have access to loops and thus have to guess at flow based on speed.
- Had difficulty in making Caltrans comfortable with PeMS data fusion.

5.4.2 CALTRANS HQ

- Any purchased data must be valid, accurate, reliable, and delivered on a timely basis.
- Concerned about PeMS data and fusion quality. Wants ability to perform QA on PeMS data. Suggests Bluetooth readers or probe data as a means to perform the QA.

- GPS data can provide 1–2 meter accuracy, and INRIX can provide this data for 70–80% of the time.
- Caltrans is satisfied with PeMS.
- Speed data will be the first data purchased from third parties.
- Third-party providers use algorithms to “smooth” data, and this results in inaccuracies in changing conditions.
- Some third parties (INRIX) don't provide historical data.
- Bluetooth is an alternative if third-party replies to RFP are inadequate.

5.4.3 MTC

- Planning on purchasing third-party data on an upcoming 511 task order. Focus is on speed data to generate travel times.
- 511 purchases speed data from SpeedInfo for $110 per unit/month.

5.4.4 CCTA

- CCTA hires consultants for new projects which allows for greater efficiency and a decrease in spending (no unions and no layoffs).

5.4.5 DELCAN

- Travel times from NAVTEQ or Traffic.com
- SpeedInfo speed data

5.5 SOLUTIONS

5.5.1 BTS

- Suggestion that flow data may be better for prediction of travel times

5.5.2 CALTRANS HQ

- Bluetooth technology is the best way to provide future travel times.
- Bluetooth is an alternative if third-party replies to RFP are inadequate.

5.5.3 CALTRANS DISTRICT 7

- The combination of probe data and historical data could work to allow for an educated guess at incident and congestion detection.
- It is possible that probe data could play some role if real-time calculations can be built in conjunction with historical data, but more computing power would be needed to store all this data.
- Probe data could also be used as a data trigger to validate scenarios without volume/occupancy data.

5.5.4 MTC

- 511 interested in exploring Bluetooth technology as an option for the probe data.

5.5.5 CCTA

- Loop detectors should be moved further out for ramp metering and properly coordinated with the mainline.
- Vehicle integration and connected car technologies look promising.

5.5.6 DELCAN

- Combine data sources to effectively close gaps by using video detectors, radar detectors, and probe data.
- California ATMS would be enhanced by increased probe data because the lack of reliable data is currently hindering systems.

6 ANALYSIS OF PREVIOUS REQUESTS FOR PROPOSALS

To see what could be learned from past experience, we reviewed the following procurement efforts:

- Caltrans' Invitation for Bids—No contract awarded
- Missouri DOT's RFP—No contract awarded
- I-95 Coalitions RFP—Contract awarded
- Wisconsin DOT's RFP—Contract awarded

6.1 COMPARISON OF STATE DOT PROCUREMENTS

The following tables show a detailed comparison of provisions across the four aforementioned proposals, broken into:

- Traffic data requirements
- Data services
- Physical facilities
- Contract pricing and payments
- Proposal requirements
- Legal and licensing issues

6.1.1 TRAFFIC DATA REQUIREMENTS

Requirement	Caltrans	Missouri	I-95 Coalition	Wisconsin
Data elements	Lane by lane for speed, volume, and occupancy	Traffic counts, average speeds, vehicular classifications (at least 3 classes), and occupancy per lane	Mean travel time and speed (lane unspecified). Desired or optional: occupancy, volume, event/incident data, and others.	Speed, travel time, communication errors, and incident flags
Frequency/Resolution	Maximum interval: 30 seconds for all three types of data	Every 60 seconds	When mean speed changes by 3 MPH or more, travel times by 5%+, or status flag change OR at least once every 5 minutes	One minute reports

Accuracy & Quality	Speed: +/-10% of ground truth Volume: 95% of ground truth Occupancy: 95% of ground truth	-	Across 3 speed groups: Average Absolute Speed Error: +/-10 %; Speed Error Bias: +/-5 % AND Score indicating confidence of estimate AND Validation by independent contractor	Must provide a method indicating stat. significant link speeds of +/- 5 MPH and stat. significant drops of 10 MPH due to traffic or incidents.
Timeliness of reporting (latency)	Within 60 seconds of each 30 second collection period	"Made available in real-time"	Required: 8 minutes or less. Highly desired: 5 minutes or less.	Max 180 seconds (3 minutes)
Availability/Reliability	Availability: 95% of the time. Completeness: 95% of the data. Reporting: self-report when below limits.	Need to have a quality control plan to include at minimum: maintenance and operation of field detection systems and data verification	Subscription services 99% 24/7 with 40 hours per year allowed for maintenance; 95% of reporting intervals should be reporting estimates.	Not explicit
Coverage	Detection locations specified	3 specified freeway sections	Map specified, deviations allowed	5 inter-city corridors
Data manipulation	Prohibited OR Detailed description of algorithm	-	Imputation prohibited	-

6.1.2 DATA SERVICES

Requirement	Caltrans	Missouri	I-95 Coalition	Wisconsin
Format of data delivery	Format should be consistent; if changing format must alert Caltrans. Must report field elements if possible	"Acceptable format" according to executed contract	Required: XML via web-based subscription service, minimum 40 subscriptions with ability to scale up to 200	XML Data stream
Monitoring service	-	-	Web-based monitoring system required (not public access)	-
Archiving services	Buffering/short term archive required	Offeror's data warehouse to store 5 min, 15 min, 60 min, and 24 hour historical	Required	-
Interface with existing systems	Contractor responsible for all data integration, including transmittal. As long as format is consistent, it can be incorporated into PeMS.	Offeror responsible for interfacing with existing systems, including costs	Must comply with TMDD standards or other open/published standards.	-

6.1.3 PHYSICAL FACILITIES

Requirement	Caltrans	Missouri	I-95 Coalition	Wisconsin
Access to DOT facilities	Permitted on case-by-case basis. Bidders proposed using loop detectors and magnetometers.	-	Access prohibited.	-

Physical facilities	-	If using field detection stations, must be every 1.5 miles at minimum and near existing interchanges when possible. Assumes they will be using physical stations.	Explicitly encouraged probe-based technologies and new and innovative approaches, including data integration.	Stated interest in mobile probes.

6.1.4 CONTRACT PRICING AND PAYMENTS

Requirement	**Caltrans**	**Missouri**	**I-95 Coalition**	**Wisconsin**
Payment frequency	Monthly	Not clear, possibly annual	Monthly	Start-up payments, then annual service fee paid in 12 monthly increments.
Payment contingencies	Initial acceptance: contingent on results of validation process. Contractor would be given time to calibrate or correct. Ongoing: for any detection site that provides traffic data meeting technical specifications in contract.	-	Determined by a formula that is a function of uptime and reliability. Right to renegotiate coverage, costs, and/or requirements or terminate contract based on results of validation.	Whatever fraction of service disruption will be reduced from monthly invoice.
Pricing model	Individual detection locations and project total costs: monthly costs and annual costs for year 1-3	Required deployment and annual operation costs	Required with both startup costs and subscription fees	Required startup cost and annual service fee, separated by mobile phone and fleet based solutions
Pricing rules	Required to report pricing declines and extend to Caltrans	Fees submitted in proposal applicable for 1 year after signing.	Must be fixed or discounted for duration of contract. Not held to quantity estimates.	Must hold for whole contract period.

Contract Period	1 year with annual option to extend for additional year, up to 3 years	5 years, option for 5 one year renewals, can be canceled at any time with written notice	3 years	2 years with 3 possible one year renewals.
Award criteria	Meets all requirements and responsible low bidder.	Scoring rubric, preference to Missouri products when quality same or better and price same or lower. Looks like lowest.	Technical submission and Financial (pricing) submission will be submitted and evaluated separately. Technical submission will have more weight. May accept other than lowest price.	Highest scoring responsible bidder. Scores based on capabilities, staff qualifications, contract requirements, additional solutions, and cost.

6.1.5 PROPOSAL REQUIREMENTS

Requirement	Caltrans	Missouri	I-95 Coalition	Wisconsin
Wireless carrier & 3rd party agreements	Proof of contract required, letters or testimonials not accepted	-	Copies of agreement or commitment letters required	Required evidence of agreements (terms and conditions, start/end dates, quantity of phones/carriers)
References	3 client references for services it has performed within the past three (3) years that are similar in size, scope, and type of service as in RFP	Identify experience providing these types of services in last 3 years. And 3 references of previous clients in last 3 years.	Must be able to confirm ability to conform with RFP	3-6 references from past 4 years similar to this solicitation. References, if contacted, must verify high satisfaction.
Financial viability		Standard: access to books, records, accounts.	Required audited documents or annual report	Letter from bank or auditor verifying financial stability.

Narratives	Written narrative of traffic data collected and estimating procedures. Witten narrative of experience.	Written narrative of data collection, quality control plan, and proposed method of interfacing with existing database.	Detailed description of how proposal meets the requirements.	Explanation of project needs and how will exceed mandatory requirements.
Site visit	-	Standard: access to facilities	Reserved right to perform site inspections at any reasonable time	-
Privacy policy	Must be in compliance with CA IT Security Policy, have a public privacy policy, and fulfill US Privacy Laws.	-	-	Included in 3rd party agreements
Warranty	-	-	-	Performance bond equal to the value of the contract

6.1.6 LEGAL AND LICENSING ISSUES

Requirement	Caltrans	Missouri
Data ownership and licensing	Traffic data received shall be sole property of the state.	Shall be defined by private partner
Company proprietary information	-	All elements of proposal will be publicly available, so offeror be aware.

6.2 CALTRANS' INVITATION FOR BIDS

No contract was awarded for Caltrans' Virtual Traffic Monitoring Station Pilot project. The reasons for abandoning the project[2] were:

- No cost savings—Bids estimated costs that were 40% higher than Caltrans' estimate for its existing system ($10,763 per year per mile versus current estimated costs of $7,700)
- Little improvement or departure from existing technologies

Table 1 below contrasts requirements in Caltrans' IFB with similar provisions in the I-95 Coalition's Request for Proposals (RFP). The I-95 Coalition's RFP awarded a bid from INRIX at $750 per year per mile for the price of its core system.

Table 1: Differences in requirements that likely contributed to high estimated costs for Caltrans' IFB

Requirement	Caltrans	I-95 Coalition
Access to DOT facilities	Permitted on case-by-case basis. Bidders proposed using loop detectors and magnetometers.	No. Explicitly encouraged probe-based technologies and new and innovative approaches.
Data required	Lane by lane for speed, volume, and occupancy	Mean travel time and speed (lane unspecified). Desired or optional: occupancy, volume, event/incident data, and others.
Frequency/Resolution	Maximum interval: 30 seconds for all three types of data	When mean speed changes by 3 MPH or more, travel times by 5%+, or status flag change OR At least once every 5 minutes
Accuracy for volume and occupancy	95% ground truth	Did not require volume and occupancy
Timeliness of Reporting	Within 60 seconds of each 30 second collection period	Required: 8 minutes or less. Highly desired: 5 minutes or less.

In essence, Caltrans' provisions were too stringent, increasing bid estimates. An analysis of the Invitation for Bids (IFB) revealed that instead of emphasizing probe data, the IFB effectively asked companies to recreate the existing system, which is heavily focused on right-of-way sensors. We considered this seriously in fashioning our own RFP for procuring probe data (see Chapter 4, *Procuring Traffic Data*).

---

[2] Report to the Legislature Virtual Traffic Monitoring Station Pilot Status Report. April 2009. Retrieved from: http://www.dot.ca.gov/docs/reports/Report_VirtualTrafficMonitoringStationPilotReport_ACC.pdf

Chapter 3

Choosing a Site and Establishing Ground Truth

In order to purchase probe data from the commercial sector, it was necessary both to identify the highway segments for which data would be procured and to get a realistic picture of the traffic conditions at those sites (the “ground truth”) to be able to meaningfully assess the purchased data. This chapter describes our efforts to select suitable stretches of highway and capture the ground truth at those sites.

1 ESTIMATING TRAVEL TIME GROUND TRUTH

Assessing the quality of third-party probe data means, in part, seeing how closely it matches the reality of traffic on the ground (the “ground truth”). This requires:

- Processing the probe data through a model to convert the point speed measurements to travel times. We use the Mobile Millennium highway velocity model.
- Establishing a reasonable estimate of ground truth, based on measured travel times of a sample of vehicles, that can serve as a benchmark or “reference state” that the probe data can be compared to

1.1 USING BLUETOOTH SENSORS TO BUILD A REFERENCE STATE

Numerous detectors, each with its own characteristics, have been developed for measuring traffic (radar, in-road sensors, pneumatic road tubes, video cameras, etc.). To build our reference state for validating probe data, we used Bluetooth sensors (Traffax BluFax Bluetooth readers) which, when placed along the road, detect turned-on Bluetooth devices in passing vehicles.

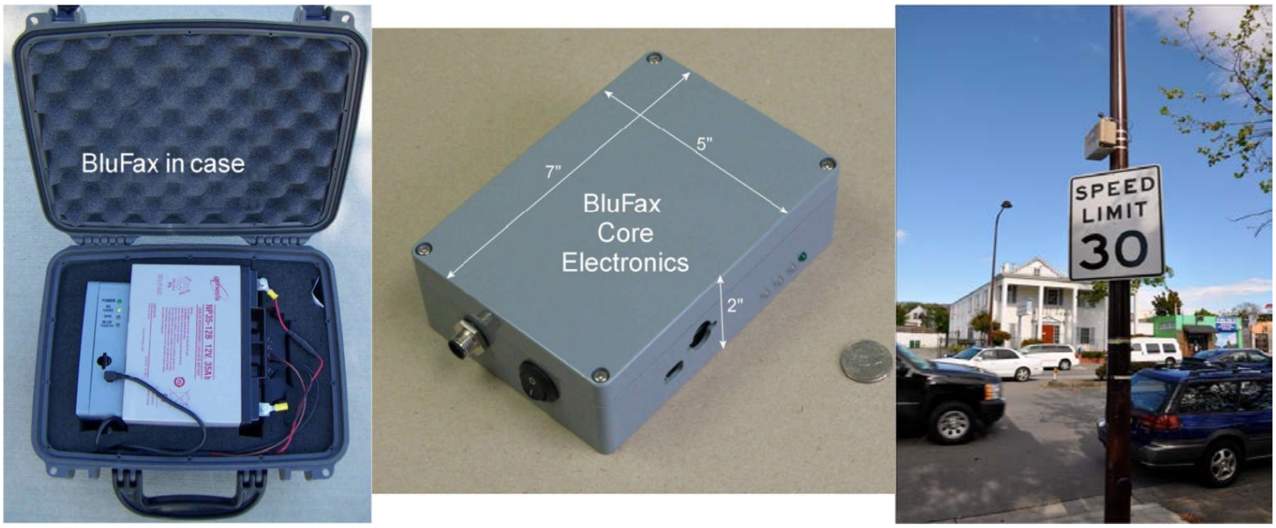


Figure 3-1: Bluetooth sensor in case and mounted on pole during the Mobile Millennium experiment

For complete sensor specifications, see section 4, *Bluetooth sensor specifications*.

1.1.1 HOW IT WORKS

When Bluetooth sensors are placed at intervals along the road, vehicle travel times are determined by capturing:

- the point in time at which a vehicle containing a discoverable Bluetooth device passes by a sensor
- the point in time at which the same Bluetooth device is detected by the next sensor

The vehicle's travel time between the sensors can then be computed. The goal is to accurately estimate the average travel times between consecutive Bluetooth sensor locations at a given time interval using the travel times of individual vehicles.

1.1.2 EXPERIMENTAL CONSIDERATIONS

Using Bluetooth technology enabled the research team to address several experimental concerns.

- **Insensitive to weather**: The sensors are weatherproof and unaffected by changes in temperature, humidity, wind, etc.
- **Self-powered**: The sensors are battery-powered units that were able to operate continuously for the two weeks needed to collect data.
- **Portable**: The sensor units are portable and small enough to be mounted on poles beside the roadway.
- **Good detection range**: With a 100-meter detection radius, the sensors were able to detect Bluetooth devices across all lanes of the highway:

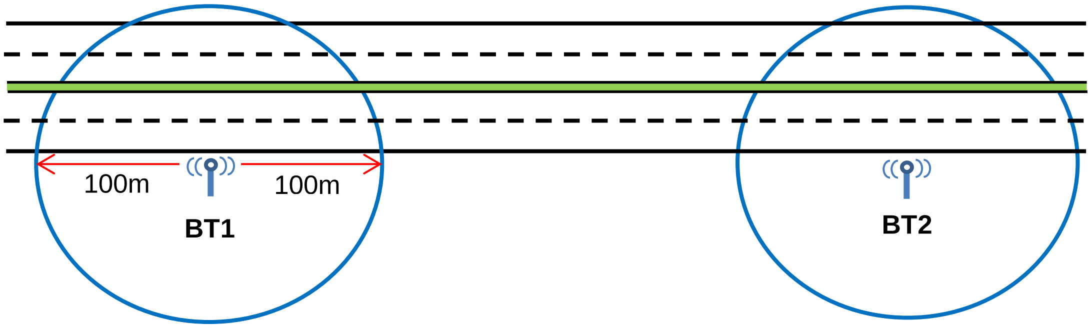


Figure 3-2: Bluetooth sensors with 100-meter detection radius

- **Acceptable penetration rate**: Since Bluetooth sensors detect all activated Bluetooth devices, we were able to detect about 3 to 5 percent of the total number of passing vehicles.

- **Re-identification**: Each detectable Bluetooth device has a unique MAC id (Media Access Control identifier) that enabled us to re-identify it throughout its path on the road network, making it possible to compute accurate travel times.

- **Remote transmission**: We received daily e-mail transmissions from the Bluetooth sensors of the previous day's recordings. That gave us ongoing confirmation that each sensor's battery had not run out and provided a quick look at the data collected the previous day.

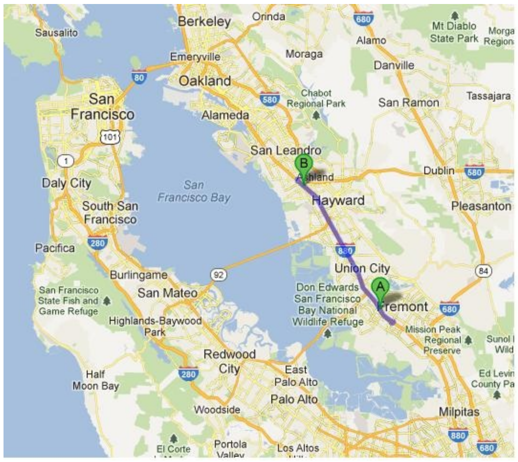
1.1.3 PROTECTING PRIVACY

To protect the privacy of Bluetooth device users, we employed a hash function that transforms the MAC id into another string which becomes the new identification of the corresponding Bluetooth device. Because there is no way of reversing the formula to get the original MAC id, people's privacy is assured.

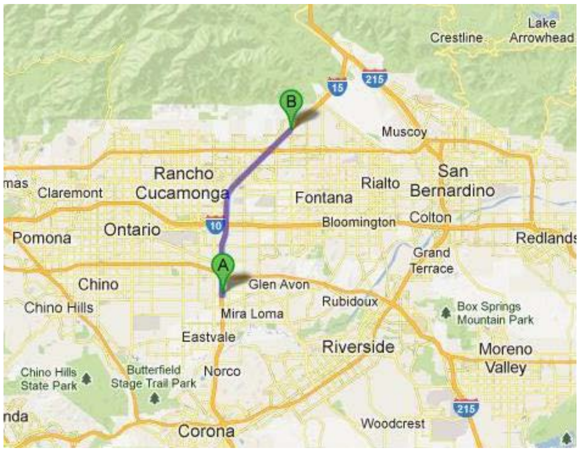
1.2 SITE SELECTION

Bluetooth sensors were deployed at three sites and detected traffic in both northbound and southbound directions at each location:

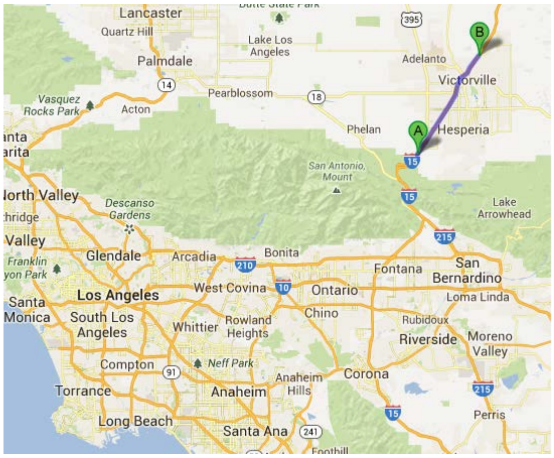
- I-880 from Fremont to Castro Valley
- I-15 in Ontario
- I-15 in Victorville



I-880



I-15 Ontario



I-15 Victorville

We were seeking locations where traffic variability was high so that:

- traffic state was not trivial to predict using historical data (i.e., it was not so obvious and predictable—such as congestion at a toll plaza during rush hour—that nothing new would be learned)
- all ranges of speeds could be assessed
- the validation could be extended to the whole spatiotemporal domain of interest

(For a description of the methodology used to study traffic variability, see *Appendix: A Study of Traffic Variability for Building a Reference State* at the end of this report.)

Using historical data available on the PeMS website, as well as historical and live data from Google Traffic, we found several segments with quite variable traffic conditions. Figure 3-3 and Figure 3-4, for example, show the evolution of speed over time for the loop detectors located on I-15 Ontario. The traffic goes in the direction of the arrow on the left of the graph (such as from sensor ID 809485 to 815332 in Figure 3-3). Each vertical line separates the days in the period studied. The average speed plotted is around 64 mph.

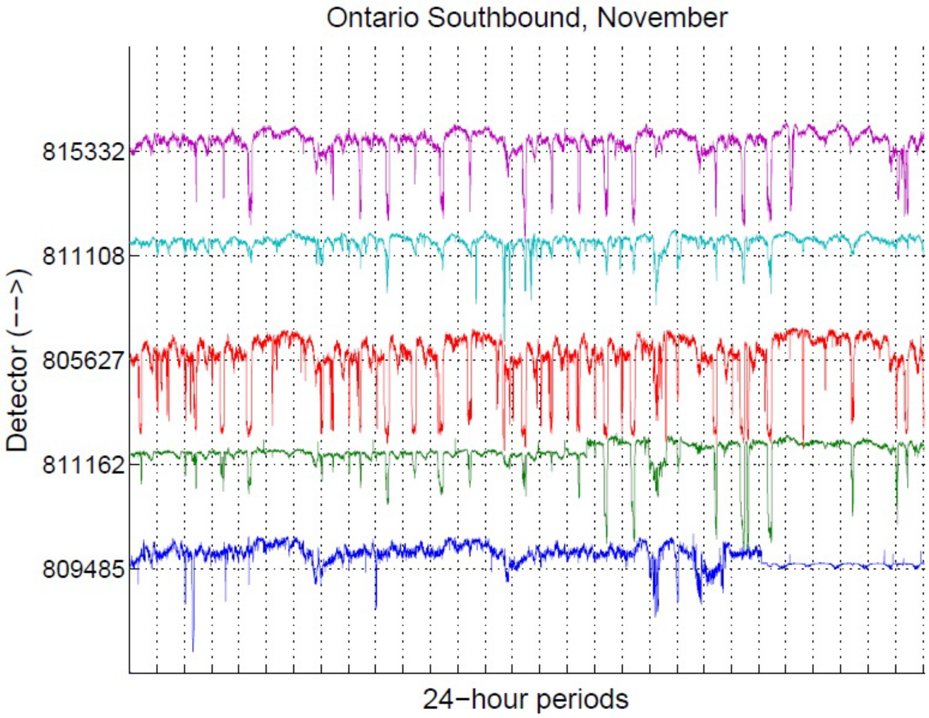


Figure 3-3: Traffic variability on I-15 Ontario southbound

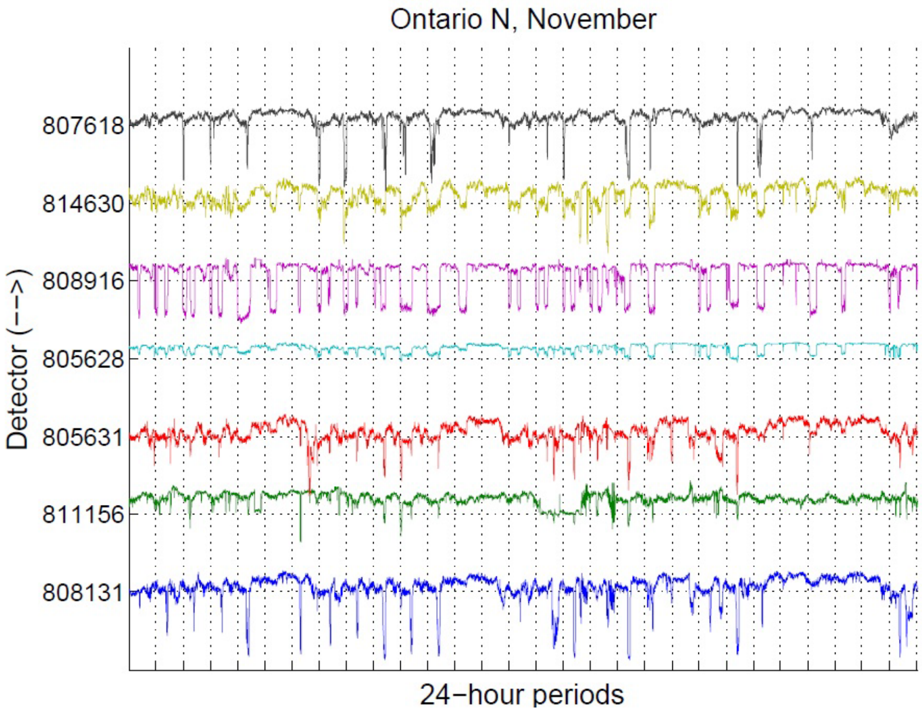


Figure 3-4: Traffic variability on I-15 Ontario northbound

The sites also needed to have lamp posts along the road on which sensors could be mounted and good lines of sight toward the highway traffic so that the antenna of the sensor had the best chance of capturing as many vehicular Bluetooth devices as possible. The research team visited the prospective sites to identify sensor placement locations and suitable installation possibilities.

1.3 SENSOR DEPLOYMENT

At each location, we deployed ten sensors for two weeks, thus gathering a total of six weeks of data.

- Individual sensors were spaced approximately 1–2 miles (1.6–3.2 kilometers) apart (see Figure 3-5). This ensured that each sensor's 100-meter range was not a significant portion of the distance between sensors.
- Sensors were installed roughly 15 feet above the road to give a direct line of sight through the windows of a car, where the Bluetooth signal is least obstructed.
- Sensors were installed away from other secondary roads, where possible, to minimize spurious readings.

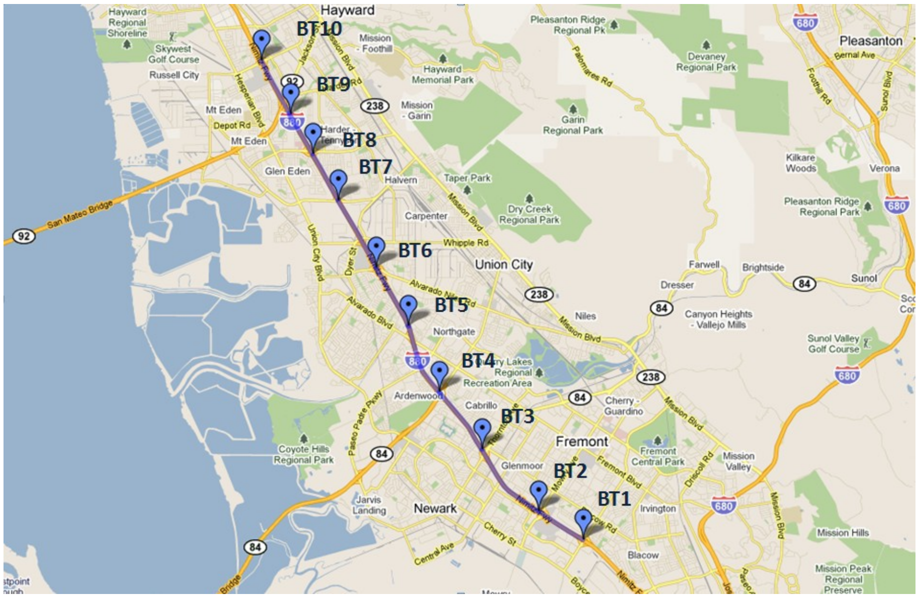


Figure 3-5: Map of Bluetooth sensors on highway I-880

For maps and photographs of all the sensor locations, see *Bluetooth sensor locations* on page 65. We are grateful to Caltrans for their kind assistance in helping position and install the sensors.

1.4 POTENTIAL SOURCES OF ERROR

The following possible errors could occur when using Bluetooth sensors:

- **Irrelevant detections**: Bluetooth sensors detect every activated Bluetooth device within range, so it is likely that a portion of the raw data obtained would be useless (for example, detecting a Bluetooth device on a nearby road that is not part of the study, or a device located on a pedestrian or bicycle).
- **Outlying behavior**: Some detected vehicles may exhibit unusual behavior on the road (such as a car stopped due to a mechanical problem), and the resulting travel times should be regarded as outliers.
- **Multiple detections**: A sensor scans for Bluetooth devices every five seconds within a 100-meter radius. This means that its range extends up to 200 meters along the roadway, and a vehicle is usually detected more than once by the same sensor as it passes through that sensor's detection field (see Figure 3-6), a problem that gets worse for vehicles traveling at low speeds or in stop-and-go traffic. Multiple readings at each Bluetooth sensor make identifying the "true" travel time a non-trivial task.

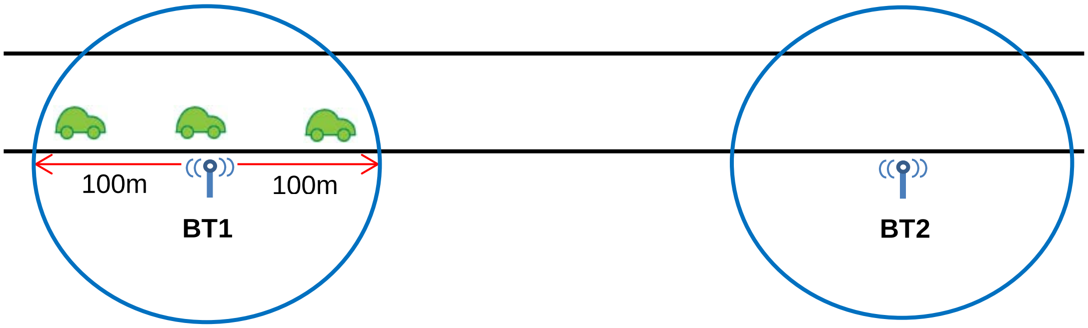


Figure 3-6: The same car detected three times (every five seconds) as it passes through sensor BT1's detection field.

- **Bias in Bluetooth-equipped vehicles**: While we are assuming there is no correlation between driver behavior and having a Bluetooth device in the car, it is possible that the behavior of Bluetooth-equipped vehicles might not accurately represent the entire vehicle population on a road segment, thus resulting in aggregated travel times that are unreliable. We believe, however, that any such bias would be negligible.

All these issues raise the need for an appropriate data filter.

1.5 FILTERING THE BLUETOOTH DATA

To address potential sources of error in the Bluetooth data, we developed a filtering process.

1.5.1 STATION FILTER

The first part of the Bluetooth filter, called a station filter, is an algorithm that eliminates duplicate entries for the same vehicle (multiple detections within a single trip) at a given sensor. Three basic filtering strategies are:

- **First time**: Using the first point in time a vehicle is detected by a given sensor works best for multiple readings due to off-ramps. This is important because an off-ramp can slow down traffic in the rightmost lane of the highway, and what we're really interested in is the travel-time for mainline flow along the highway.

- **Last time**: Using the last point in time a vehicle is detected by a given sensor works best for multiple readings due to on-ramps. By picking the last time, we eliminate the time spent on the ramp and get a more accurate estimate. This is especially important at metered lights.

- **Average time**: Using the average time of the measurements should give us an estimate of the time at which that vehicle was closest to the sensor if the multiple readings were simply due to congestion. This would be the strategy of choice in a case where there are no nearby ramps.

We applied the first time strategy to our initial filtering of the Bluetooth data. Further research on the station filter gave us a fuller understanding of the filtering parameters. That analysis is described in section 3.

1.5.2 TRAVEL TIME OUTLIER FILTER

The purpose of the travel time outlier filter is to filter out travel time values that are visibly too high or too low. A robust filtering technique known as the interquartile range filter is used to flag these values as outliers. The Interquartile Range (IQR) of a sequence of numerical values is mathematically defined as the difference between the 75th percentile and the 25th percentile. The IQR filter defines the lower and upper fences as follows:

- lower fence = 25th percentile - 2 * IQR
- upper fence = 75th percentile + 2 * IQR

The IQR filter applied on a normal distribution would flag as outlier a point that is more than three standard deviations greater than the mean. Indeed, for this distribution the one-sigma limit approximately equals $\frac{3}{4}$ /QR. But the IQR filter is more robust than a simple standard deviation filter for non-normal distributions because outliers have a significant impact on standard deviation calculation, whereas their impact on 25th and 75th percentiles is small and even non-existent most of the time.

The sequences of points are defined as 30-minute-period groups of prefiltered travel times. It means that each travel time is tested against the lower and upper fences of its corresponding 30-minute-interval group of travel times. Figure 3-7 shows the results after applying the IQR filter to one week of data from highway I-880 for one route between two Bluetooth sensors.

1.6 DISPLAYING THE BLUETOOTH RESULTS

To display the results of capturing, filtering, and aggregating the Bluetooth data, we developed a MATLAB plotting tool that enables us to generate four different kinds of data visualization plots:

- Travel time outlier flags (Figure 3-7)
- Average 15-minute detection rate (Figure 3-8)
- Time-space velocity diagram (Figure 3-9)
- Aggregated total travel time (Figure 3-10)

The following figures show the Bluetooth data gathered:

- for each week (weeks 1 and 2)
- at each location (I-880, I-15 Ontario, I-15 Victorville)
- in each direction (northbound and southbound)

Because the travel time outlier filter was applied to each *route* (each highway segment between two sensors) and therefore generated over 100 plots, we have included only one outlier filter plot as an example.

1.6.1 I-880 NORTHBOUND — WEEK ONE

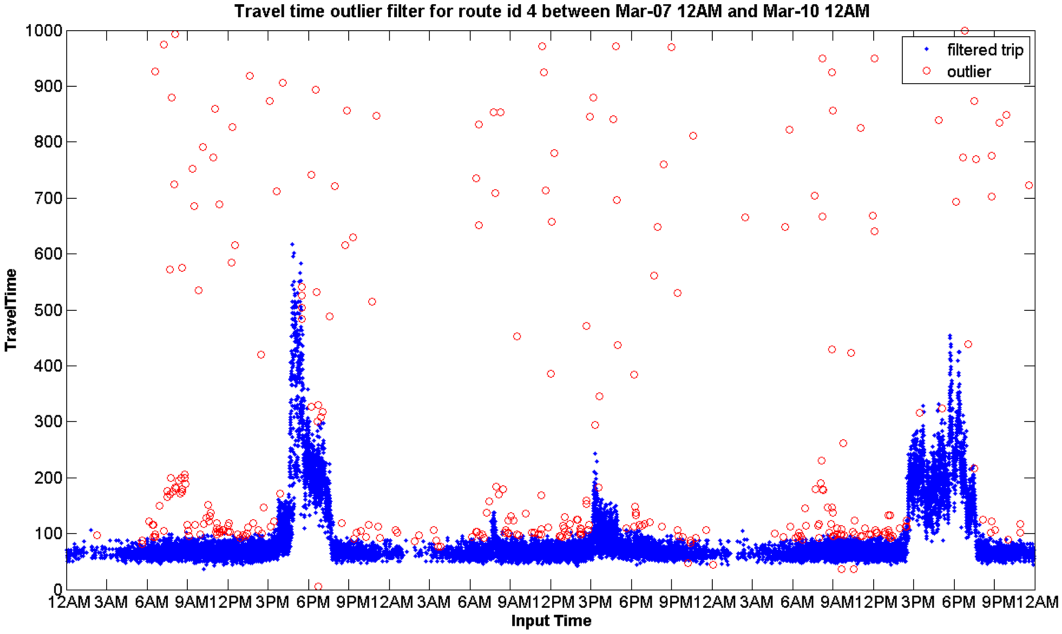


Figure 3-7: Results of travel time outlier filter for one week and one route between two sensors on highway I-880.

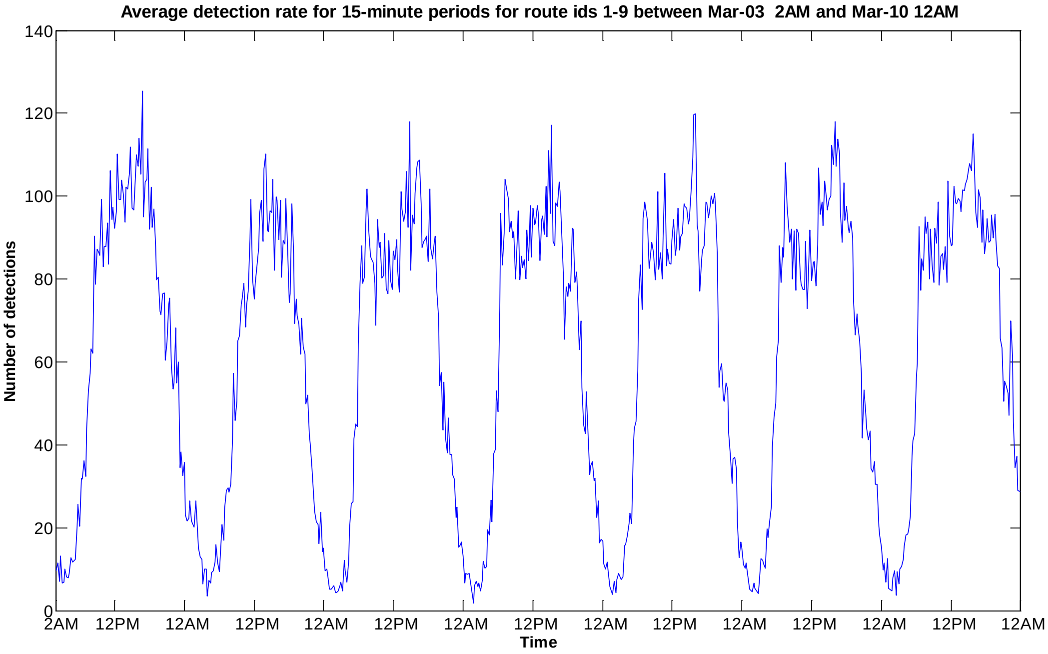


Figure 3-8: Detection rate for one week of Bluetooth data on highway I-880

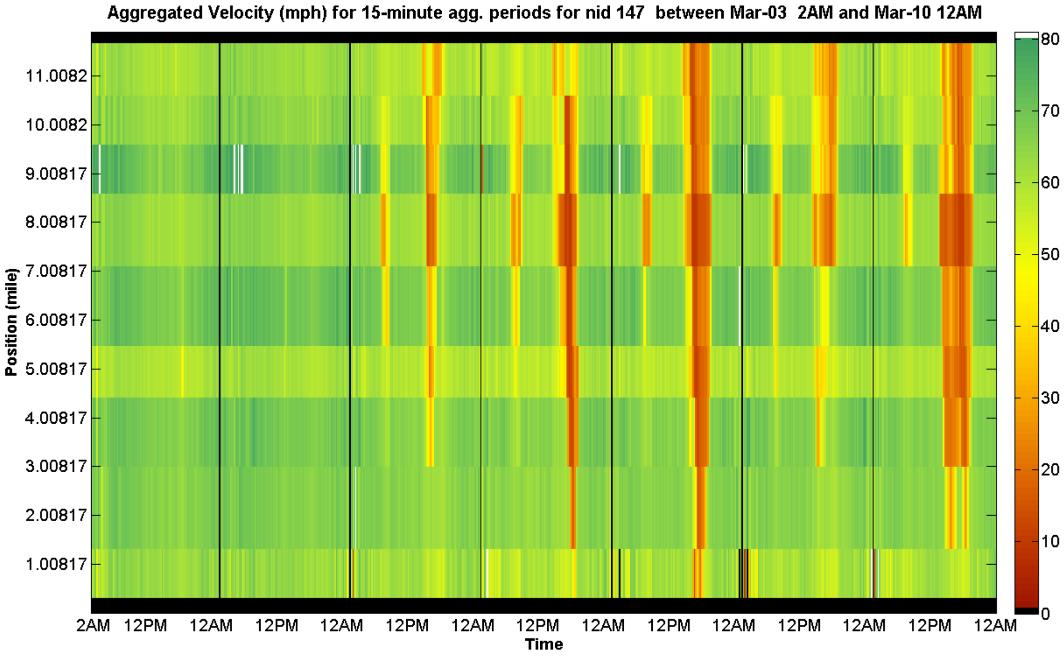


Figure 3-9: Time-space velocity grid for one week of Bluetooth data on highway I-880. Colors indicate velocity, ranging from high speed (green) to low speed (red); the scale on the right indicates miles per hour. Red shading shows traffic slowing significantly at particular locations around 5 p.m. Monday–Friday. Black vertical lines, indicating the absence of data, occur each day at 2 a.m. when all the sensors briefly shut down and restart in order to transmit the day's data. Black horizontal bands at the top and bottom of the graph (absence of data) indicate that the highway segment used in the model (the mathematical representation of the road network) is slightly longer than the actual length of the Bluetooth deployment.

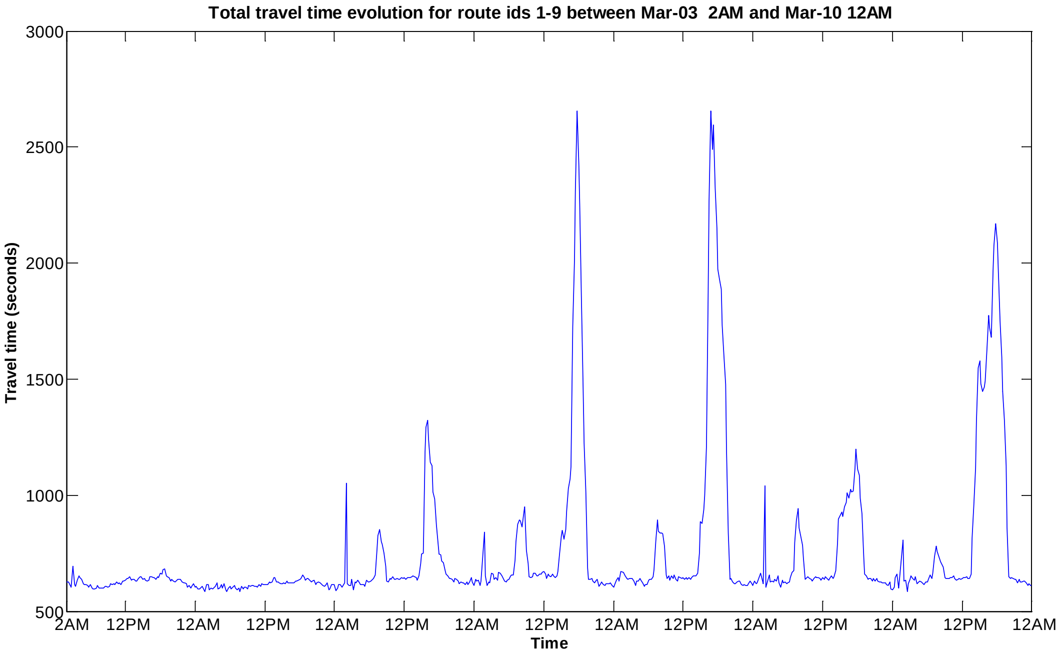


Figure 3-10: Total travel times for the same data as in Figure 3-9, showing spikes in travel time when traffic is slowest.

1.6.2 I-880 NORTHBOUND — WEEK TWO

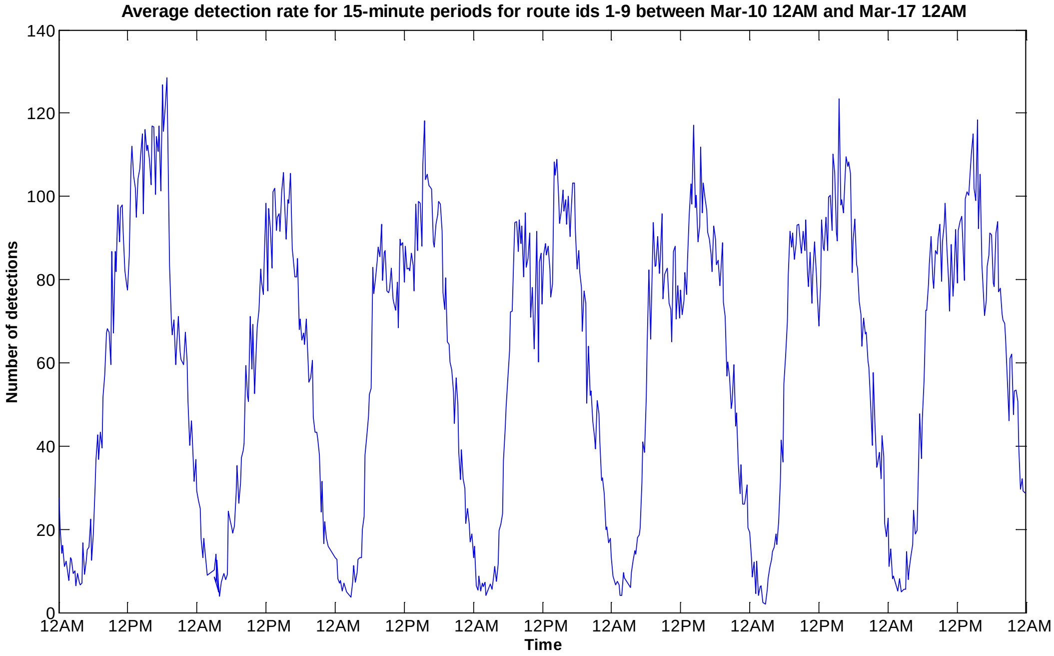


Figure 3-11: Bluetooth detection rate for highway I-880 northbound, week two

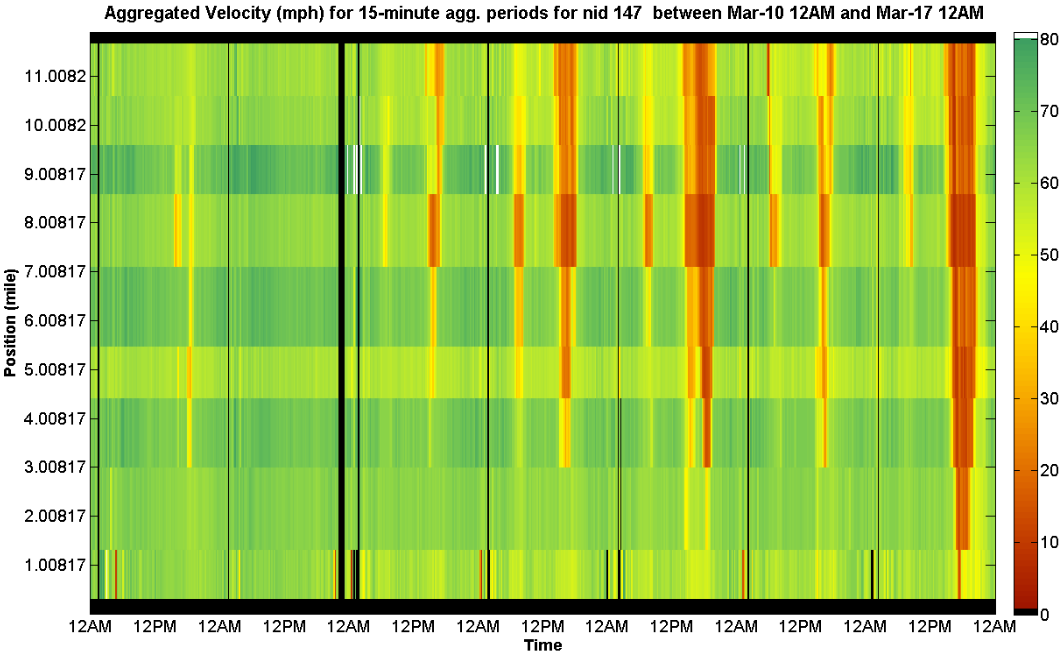


Figure 3-12: Time-space velocity grid for highway I-880 northbound, week two. The thick black vertical stripe (absence of data) occurs because of the change to Daylight Saving Time, when an hour of data appears to be lost.

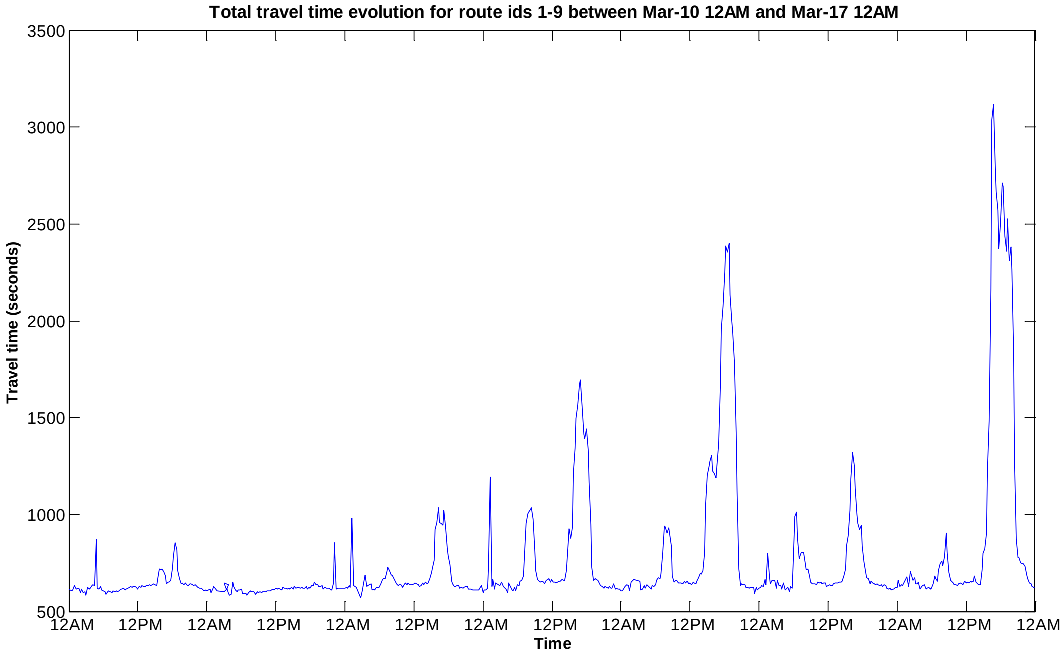


Figure 3-13: Total travel times for highway I-880 northbound, week two

1.6.3 I-880 SOUTHBOUND — WEEK ONE

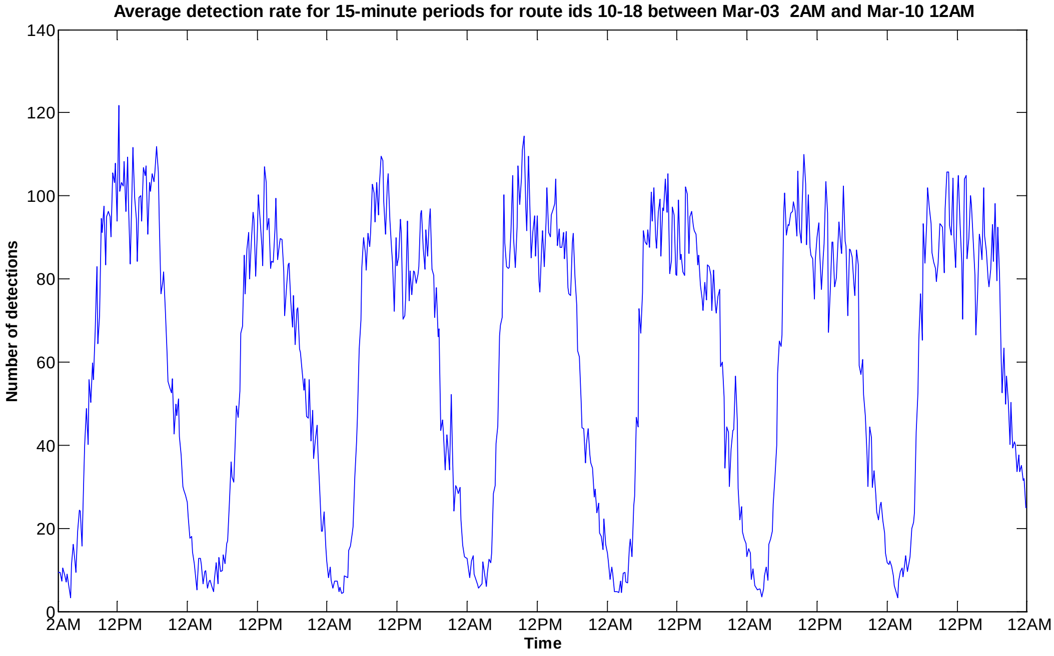


Figure 3-14: Bluetooth detection rate for highway I-880 southbound, week one

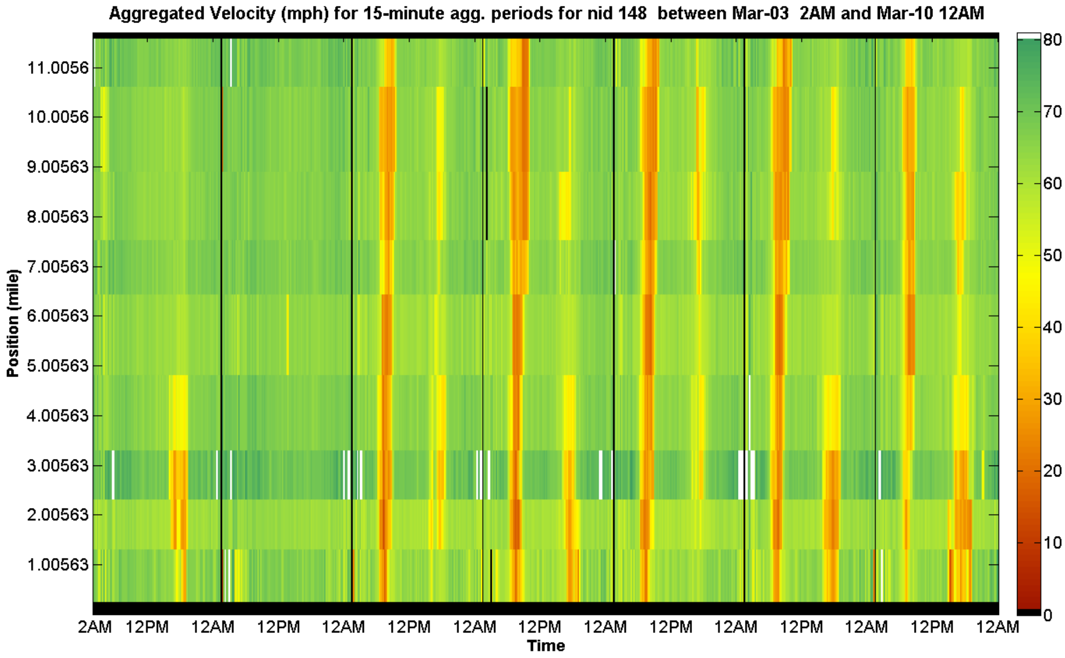


Figure 3-15: Time-space velocity grid for highway I-880 southbound, week one

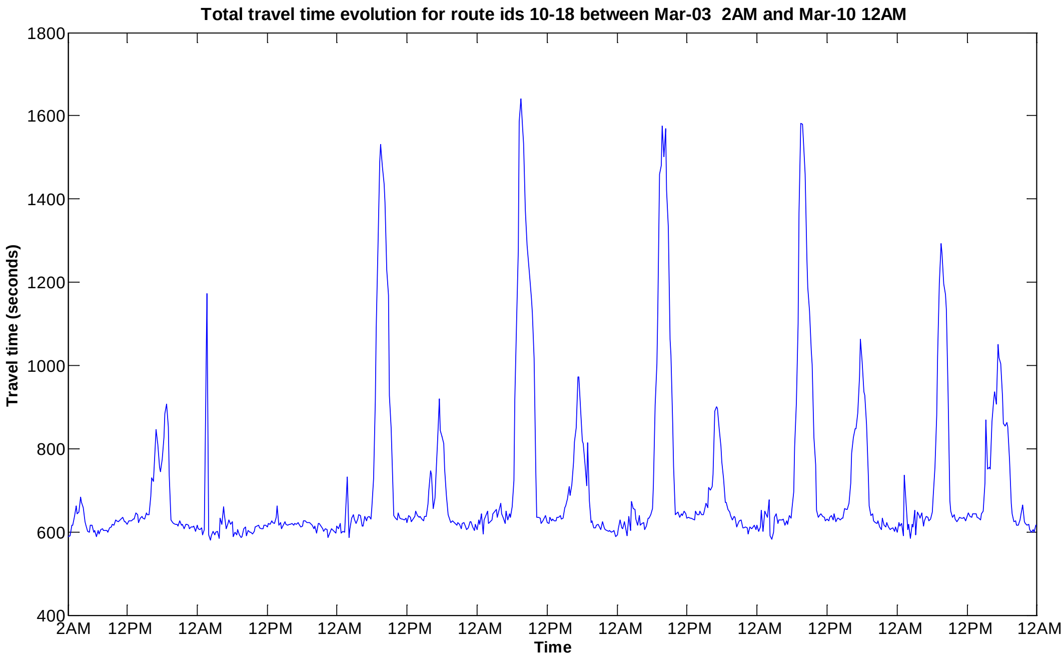


Figure 3-16: Total travel times for highway I-880 southbound, week one

1.6.4 I-880 SOUTHBOUND — WEEK TWO

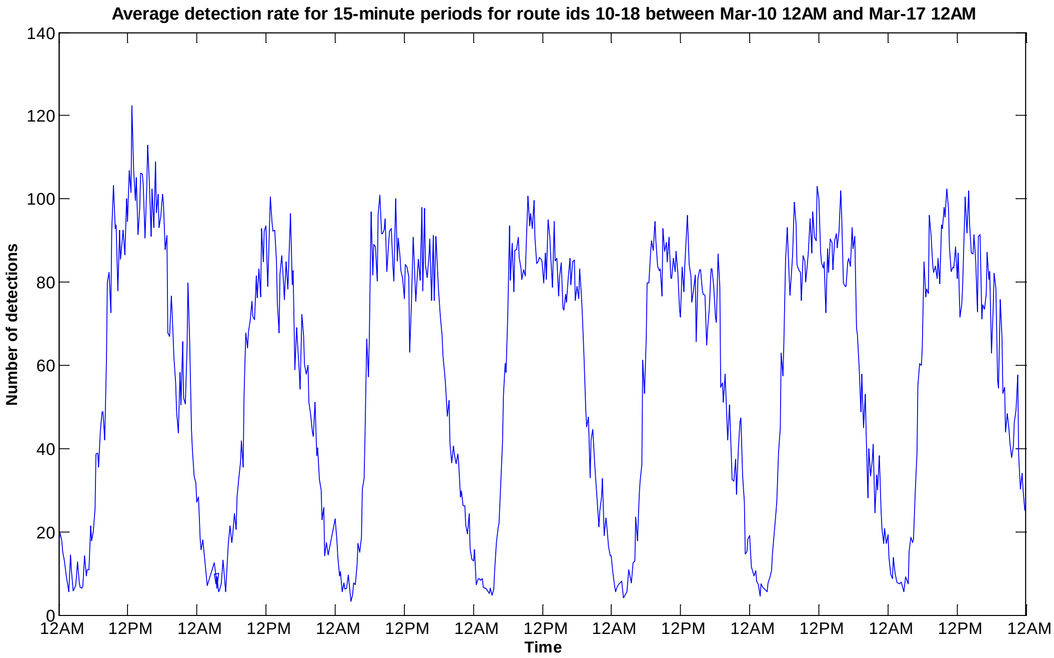


Figure 3-17: Bluetooth detection rate for highway I-880 southbound, week two

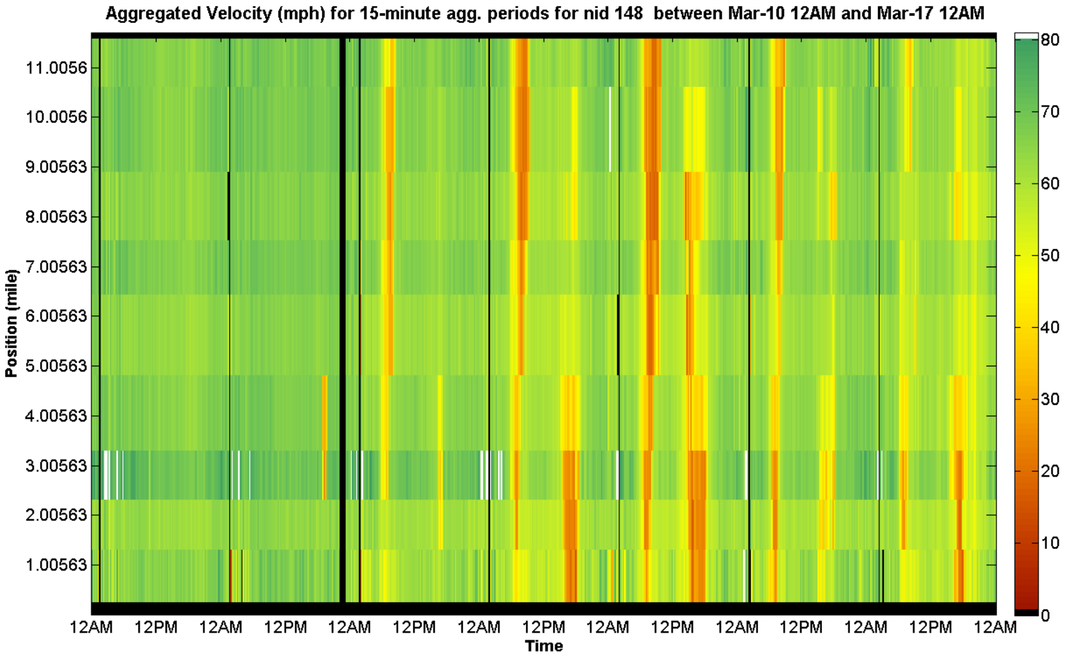


Figure 3-18: Time-space velocity grid for highway I-880 southbound, week two

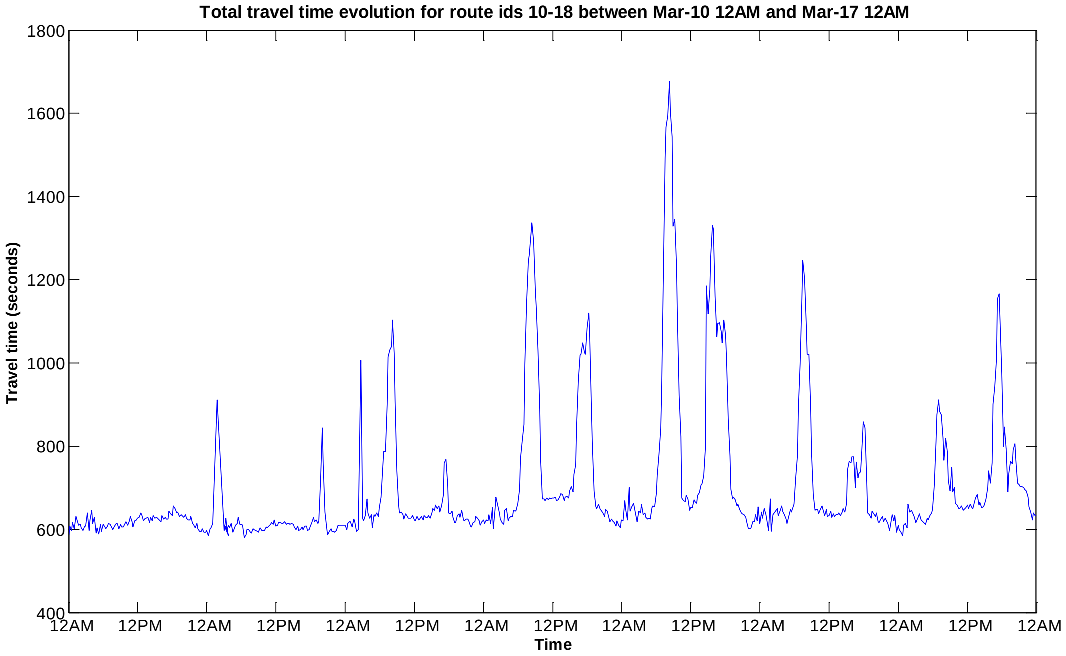


Figure 3-19: Total travel times for highway I-880 southbound, week two

1.6.5 I-15 ONTARIO NORTHBOUND — WEEK ONE

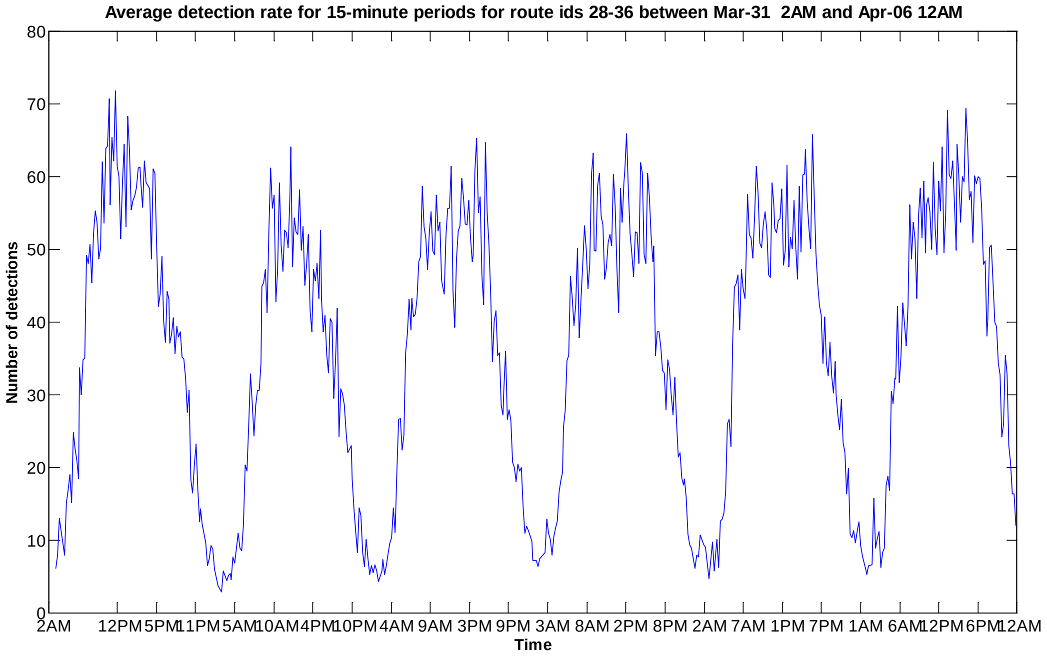


Figure 3-20: Bluetooth detection rate for highway I-15 Ontario northbound, week one

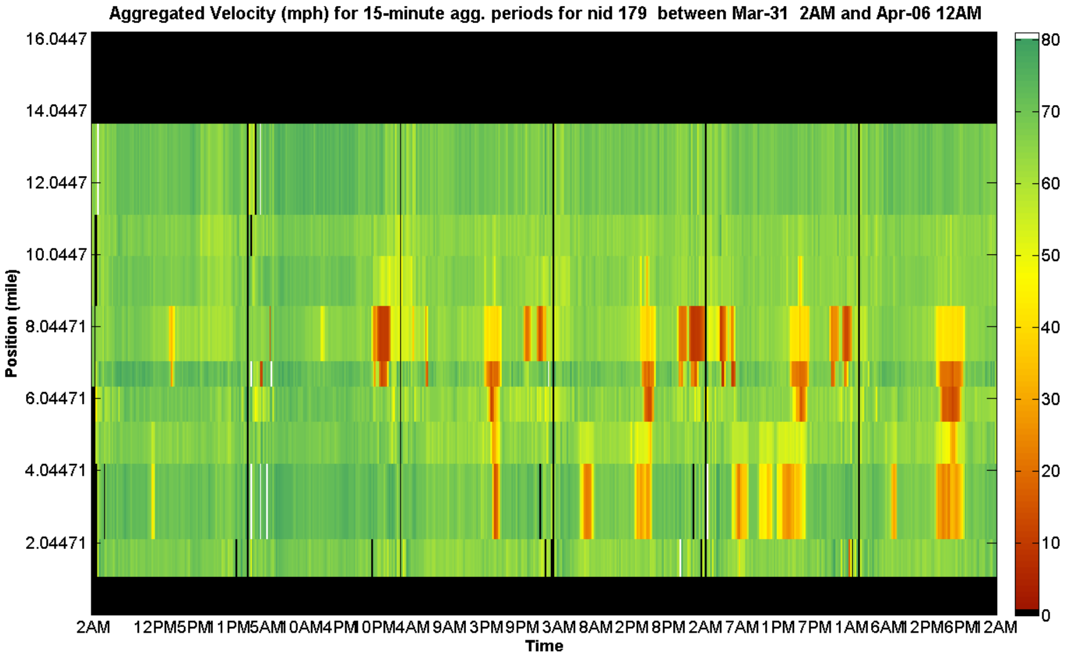


Figure 3-21: Time-space velocity grid for highway I-15 Ontario northbound, week one

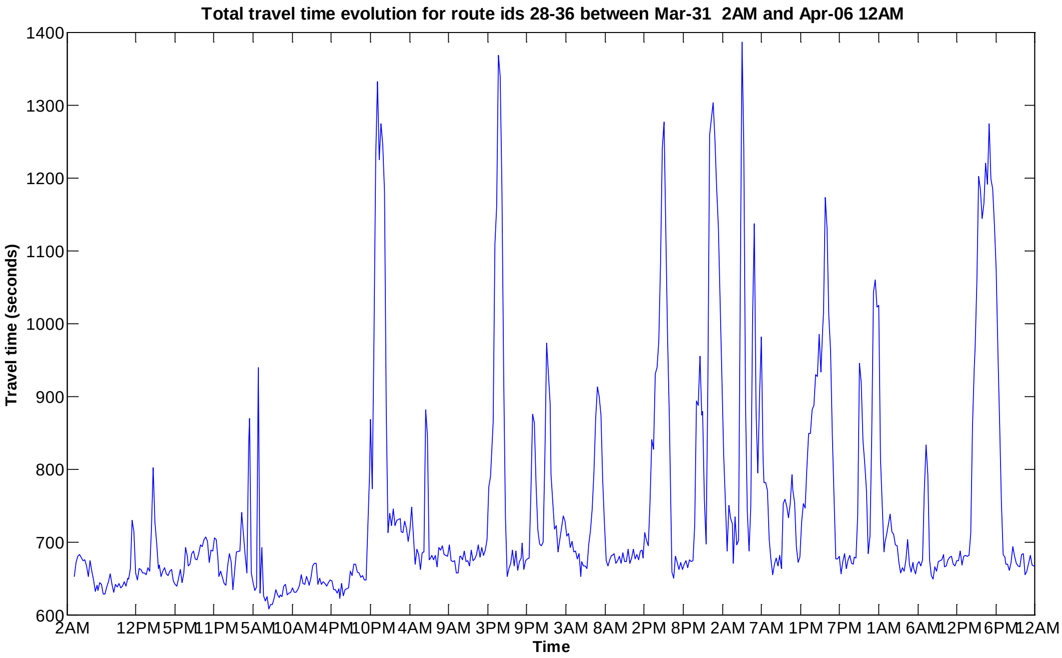


Figure 3-22: Total travel times for highway I-15 Ontario northbound, week one

1.6.6 I-15 ONTARIO NORTHBOUND — WEEK TWO

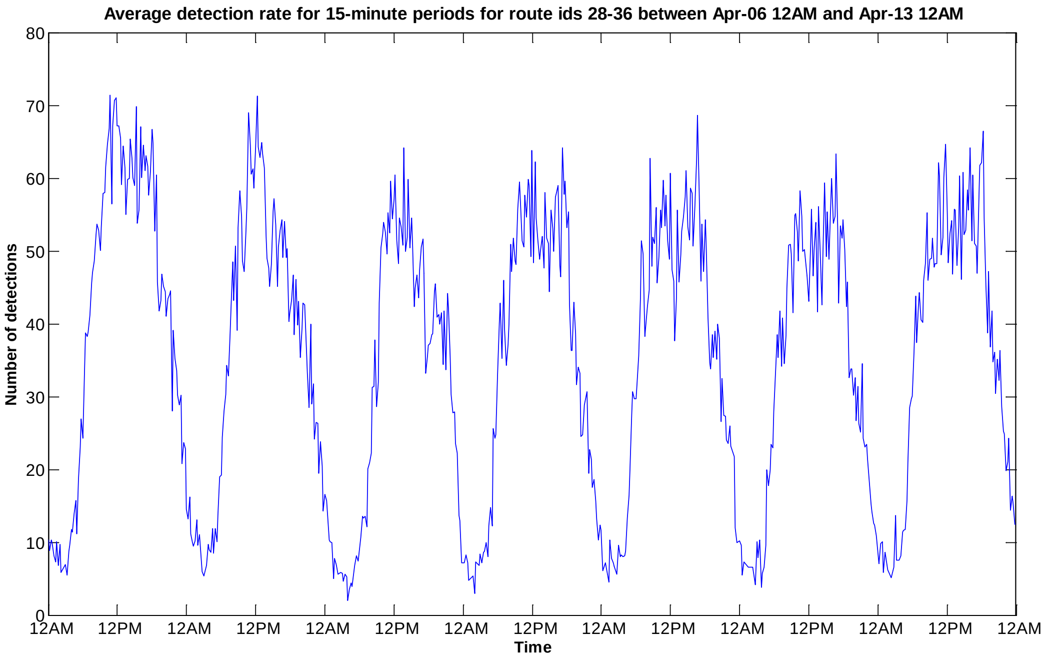


Figure 3-23: Bluetooth detection rate for highway I-15 Ontario northbound, week two

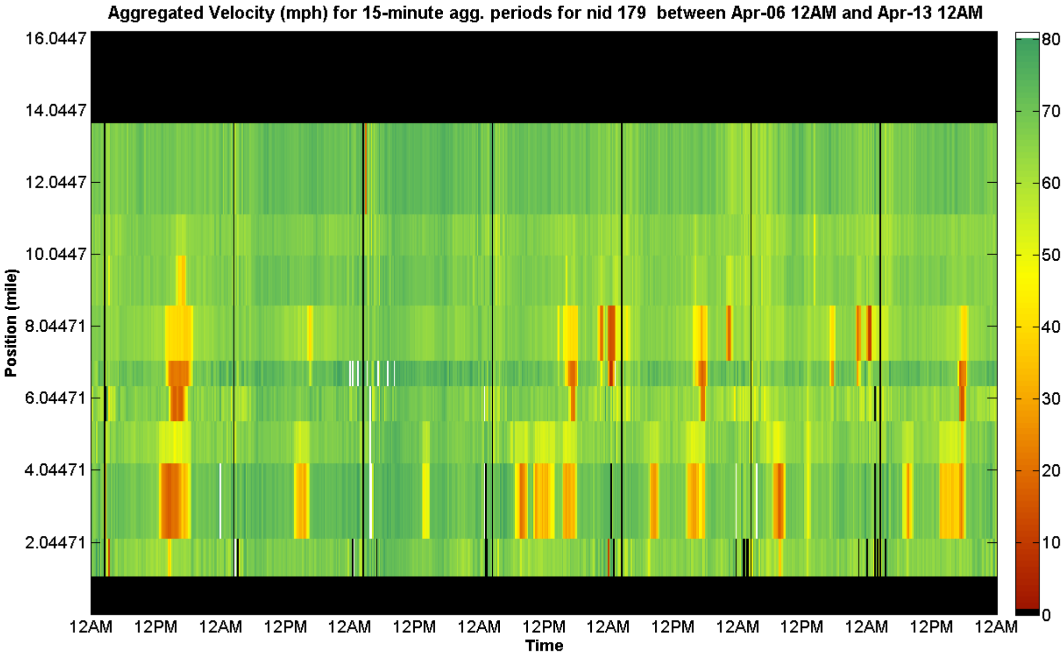


Figure 3-24: Time-space velocity grid for highway I-15 Ontario northbound, week two

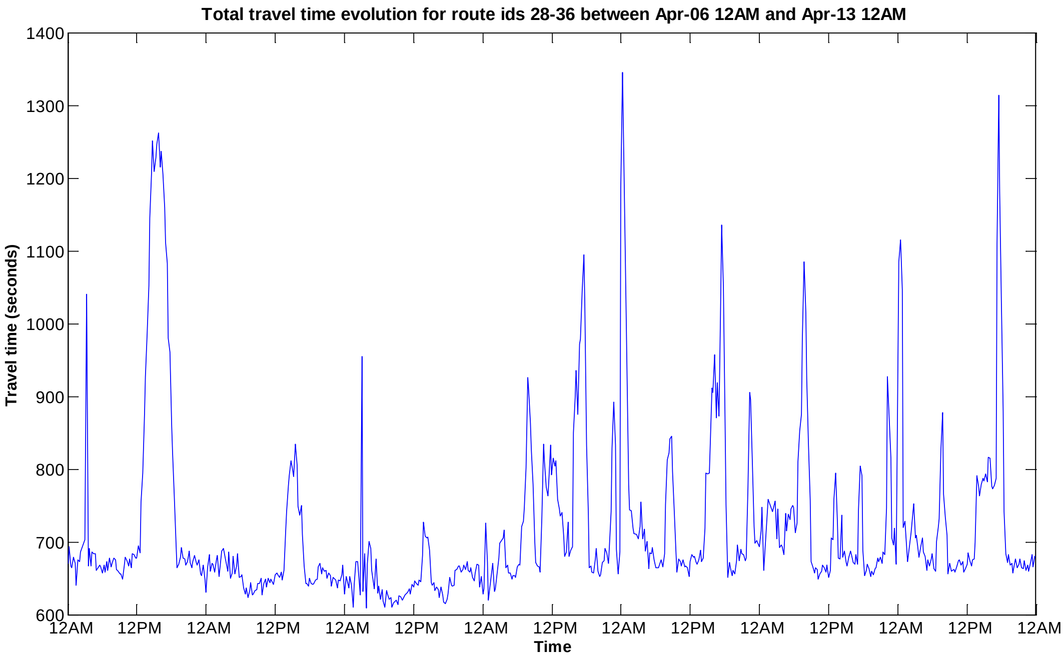


Figure 3-25: Total travel times for highway I-15 Ontario northbound, week two

1.6.7 I-15 ONTARIO SOUTHBOUND — WEEK ONE

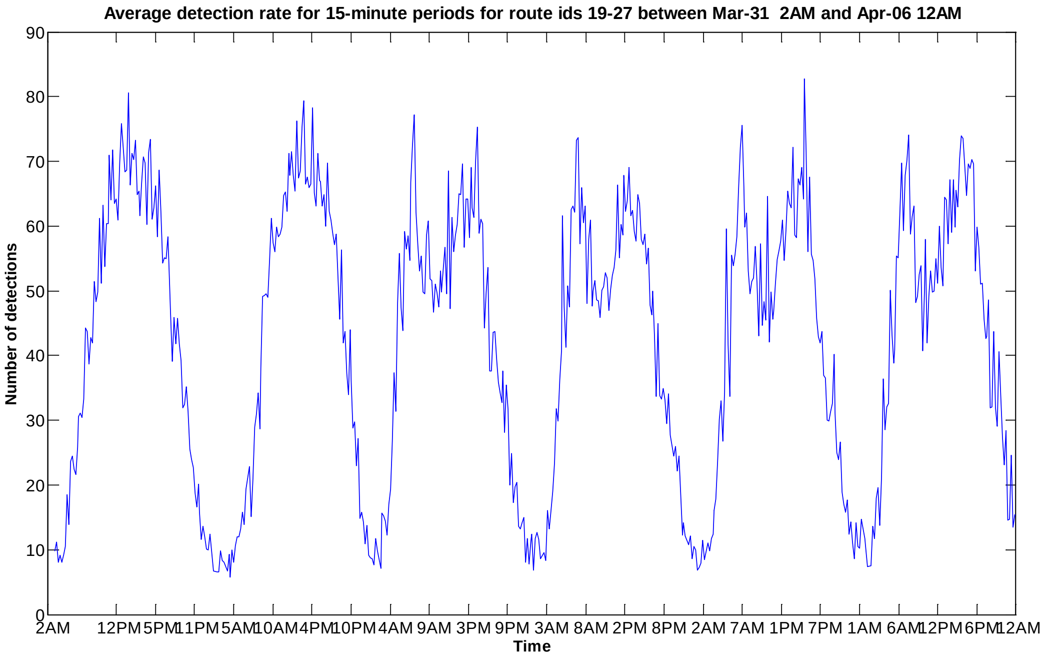


Figure 3-26: Bluetooth detection rate for highway I-15 Ontario southbound, week one

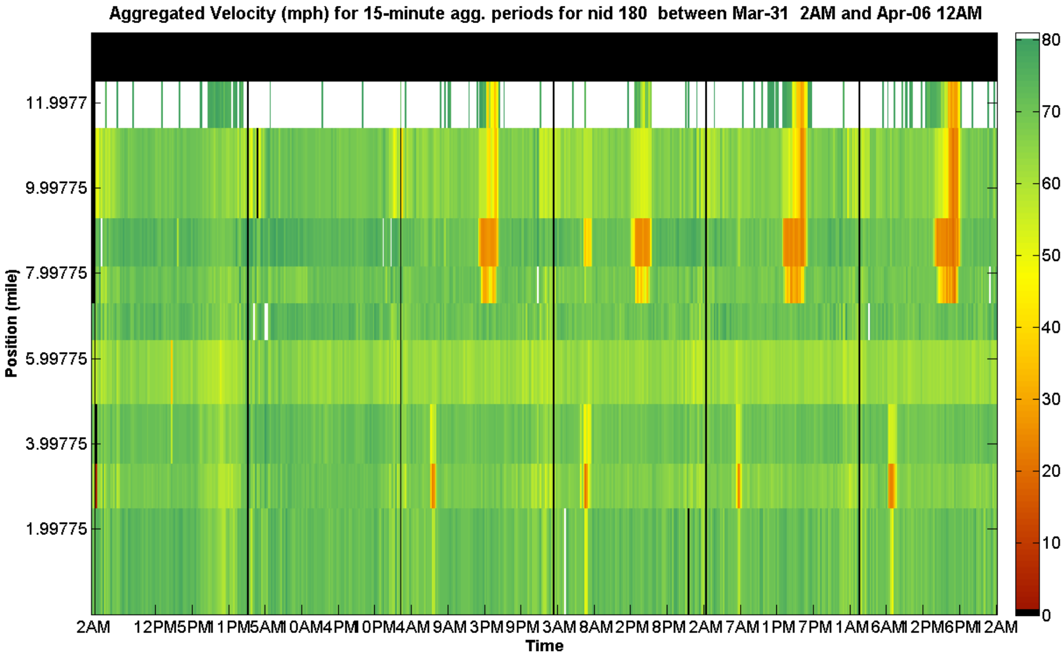


Figure 3-27: Time-space velocity grid for highway I-15 Ontario southbound, week one

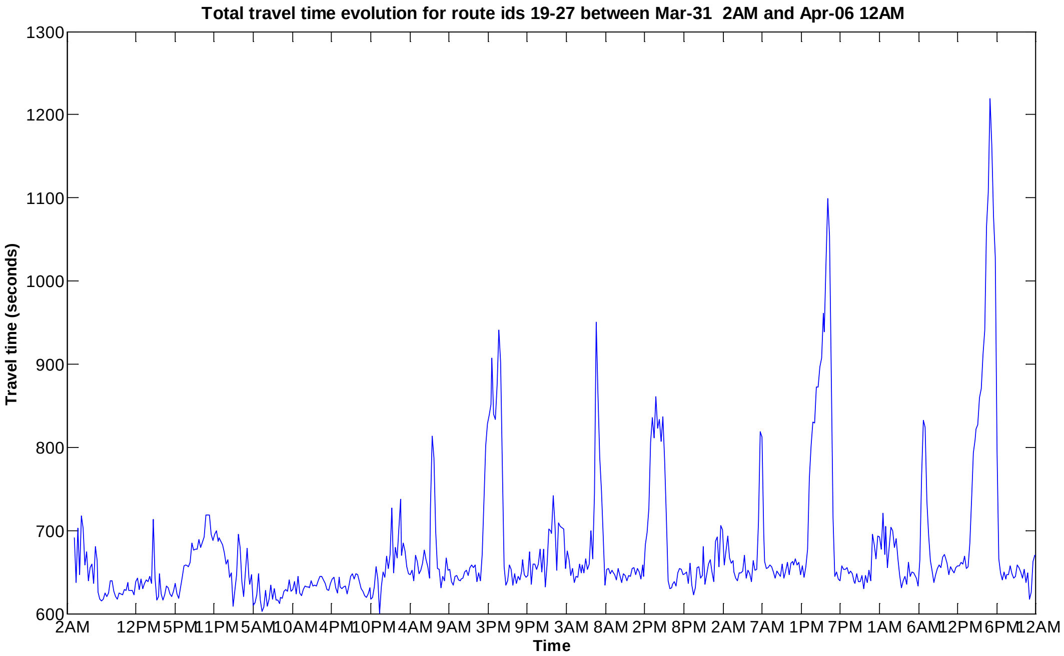


Figure 3-28: Total travel times for highway I-15 Ontario southbound, week one

1.6.8 I-15 ONTARIO SOUTHBOUND — WEEK TWO

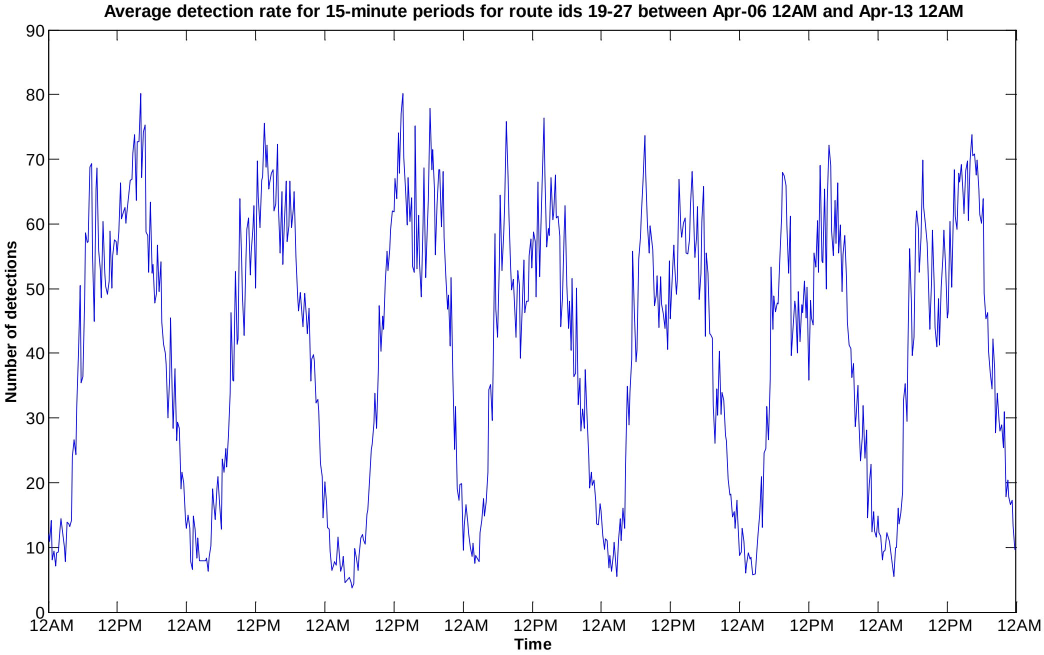


Figure 3-29: Bluetooth detection rate for highway I-15 Ontario southbound, week two

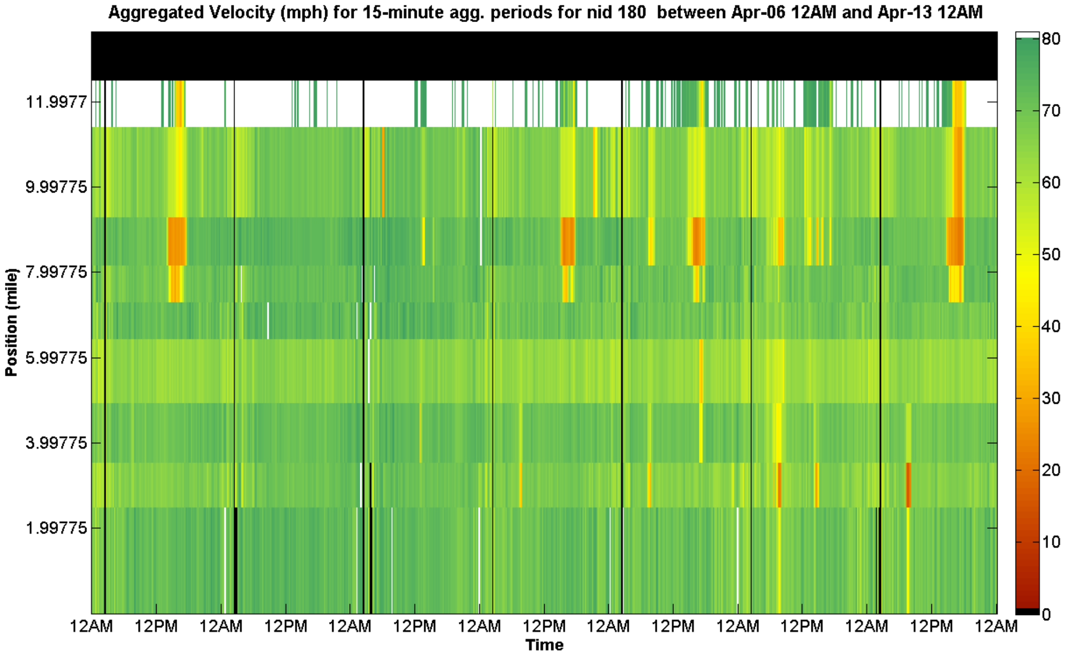


Figure 3-30: Time-space velocity grid for highway I-15 Ontario southbound, week two

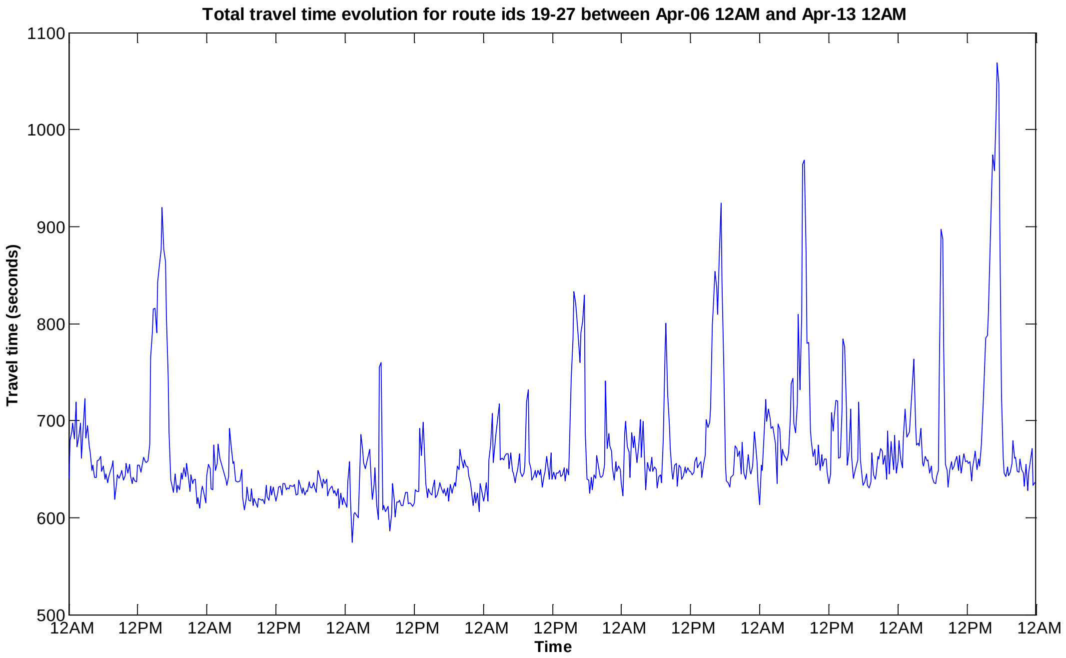


Figure 3-31: Total travel times for highway I-15 Ontario southbound, week two

1.6.9 I-15 VICTORVILLE NORTHBOUND — WEEK ONE

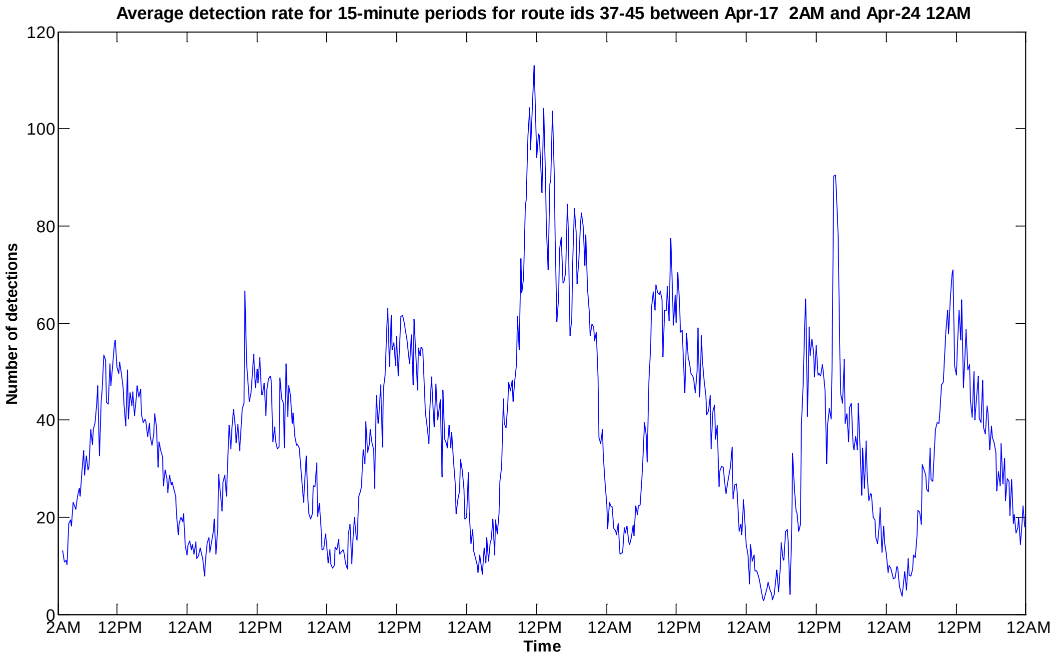


Figure 3-32: Bluetooth detection rate for highway I-15 Victorville northbound, week one

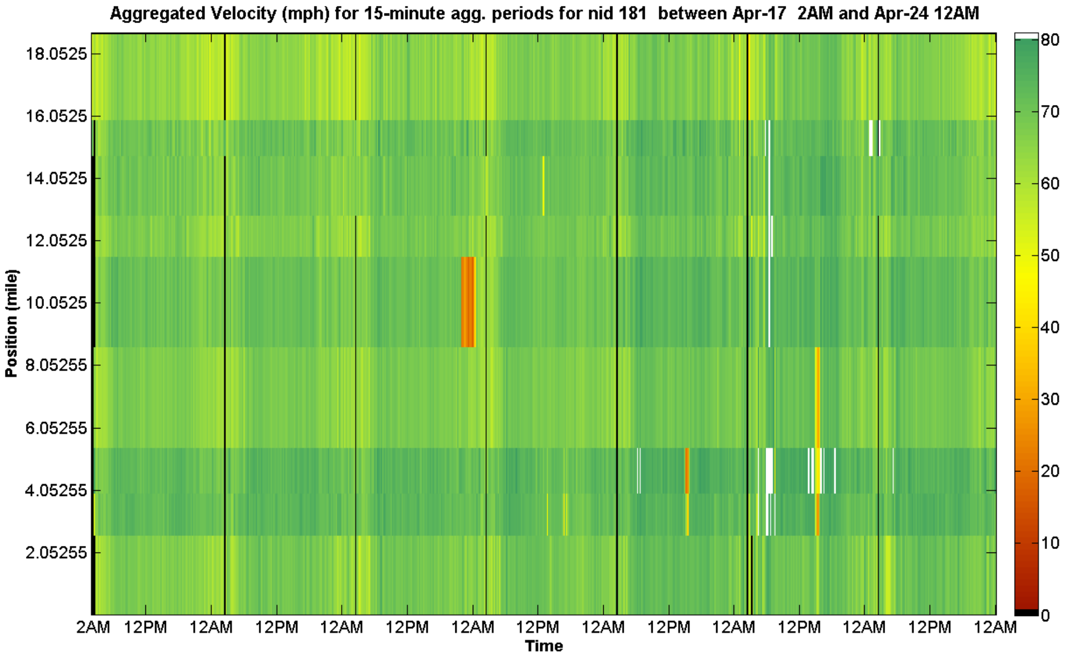


Figure 3-33: Time-space velocity grid for highway I-15 Victorville northbound, week one

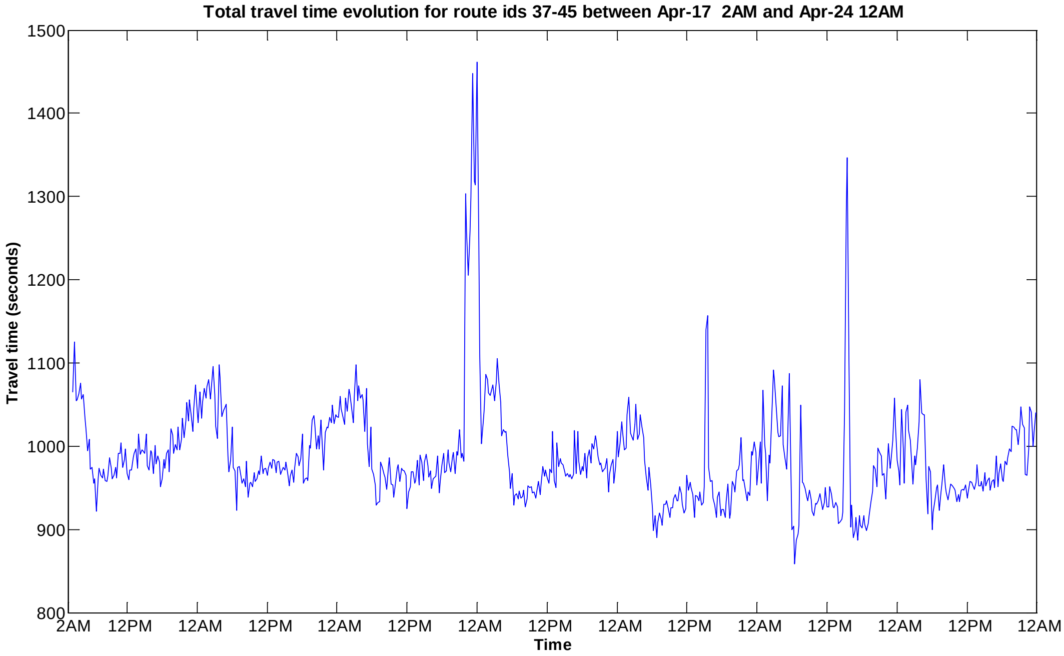


Figure 3-34: Total travel times for highway I-15 Victorville northbound, week one

1.6.10 I-15 VICTORVILLE NORTHBOUND — WEEK TWO

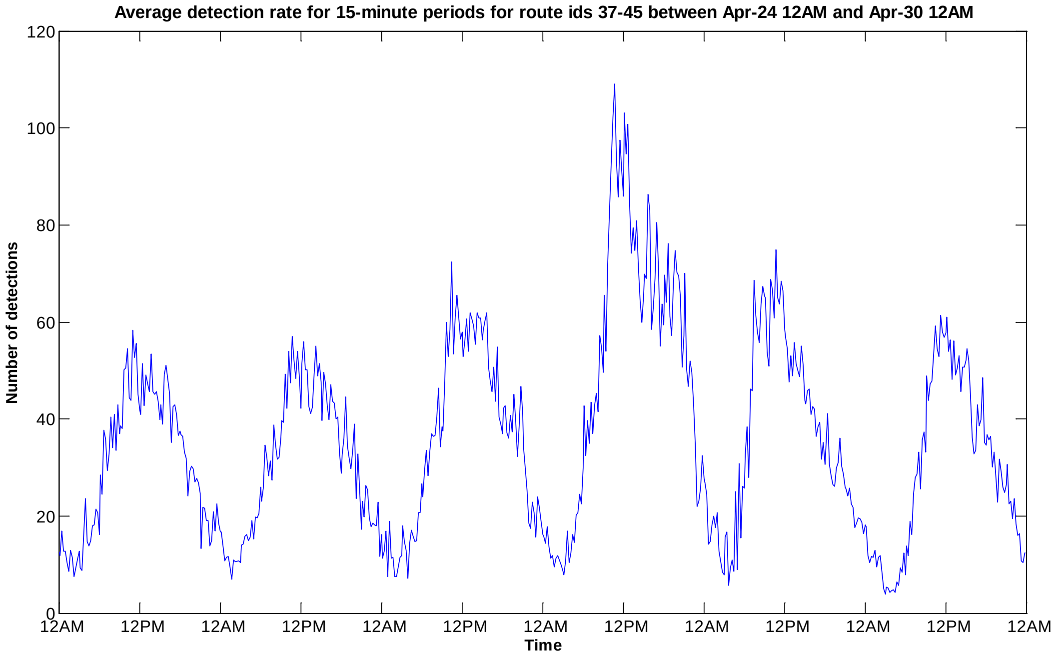


Figure 3-35: Bluetooth detection rate for highway I-15 Victorville northbound, week two

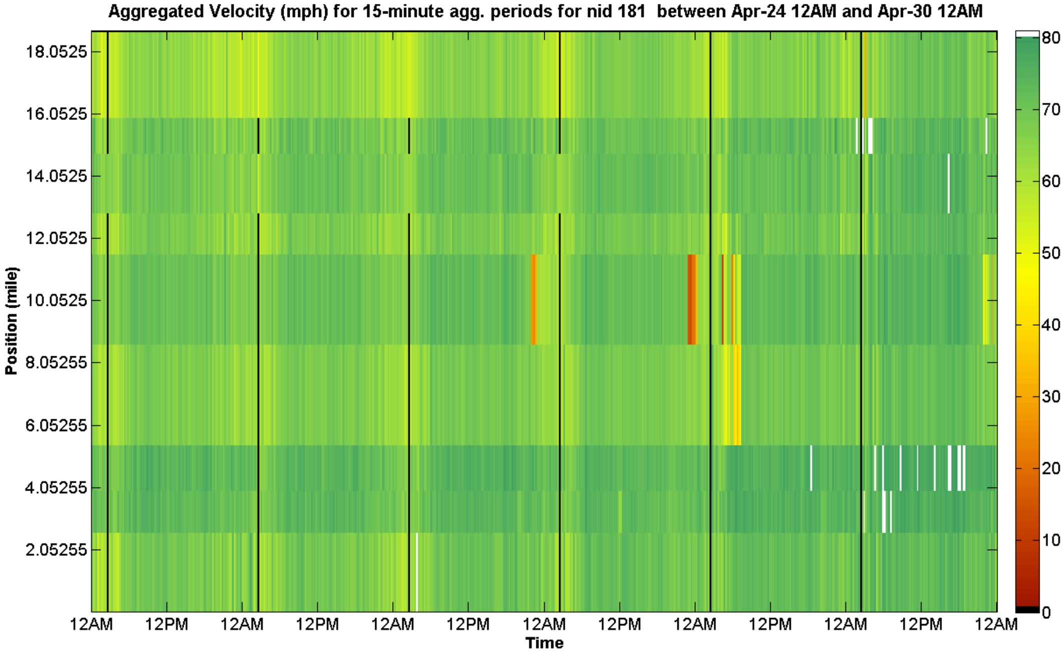


Figure 3-36: Time-space velocity grid for highway I-15 Victorville northbound, week two

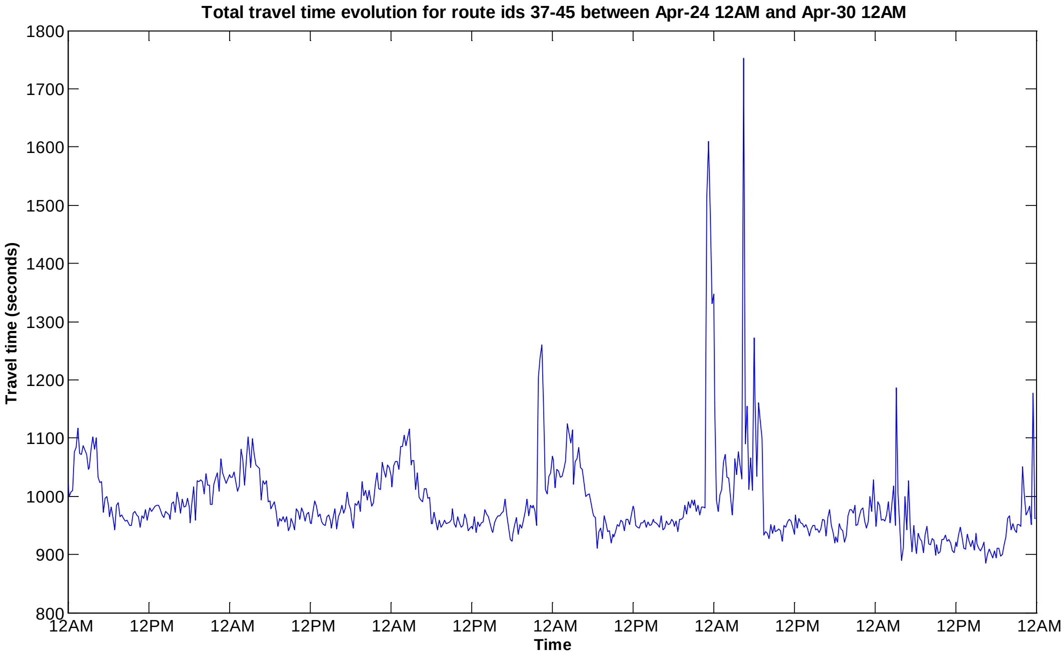


Figure 3-37: Total travel times for highway I-15 Victorville northbound, week two

1.6.11 I-15 VICTORVILLE SOUTHBOUND — WEEK ONE

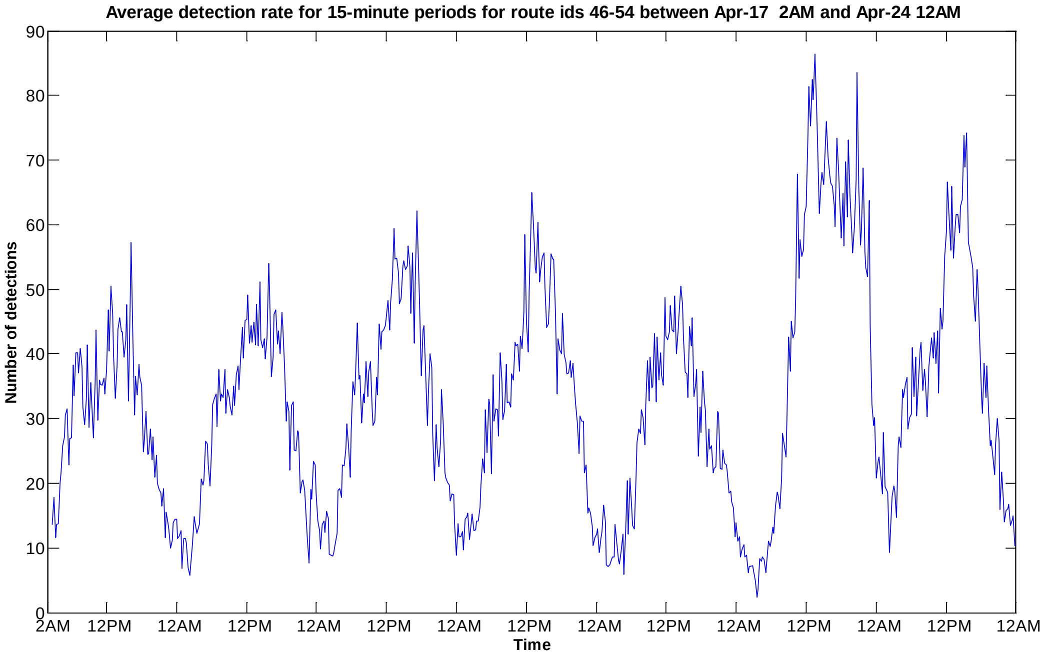


Figure 3-38: Bluetooth detection rate for highway I-15 Victorville southbound, week one

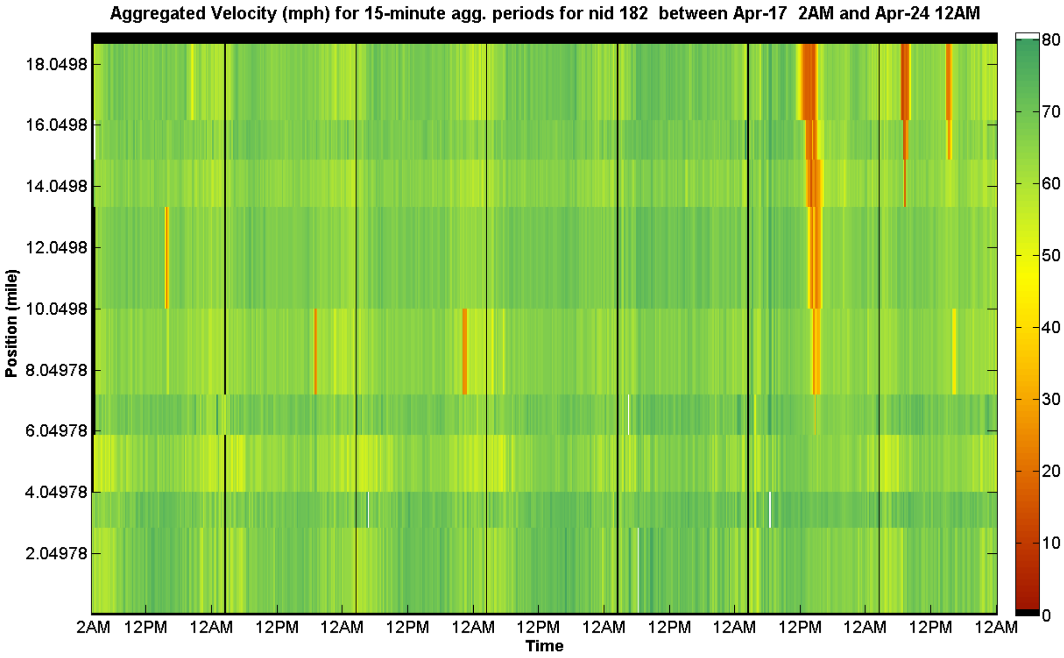


Figure 3-39: Time-space velocity grid for highway I-15 Victorville southbound, week one

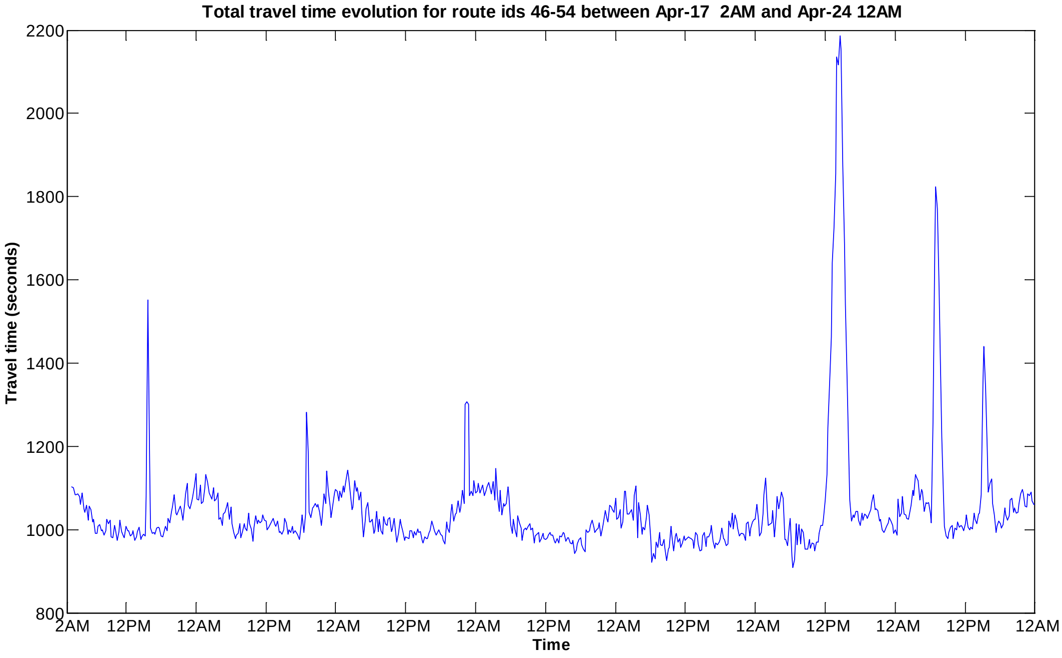


Figure 3-40: Total travel times for highway I-15 Victorville southbound, week one

1.6.12 I-15 VICTORVILLE SOUTHBOUND — WEEK TWO

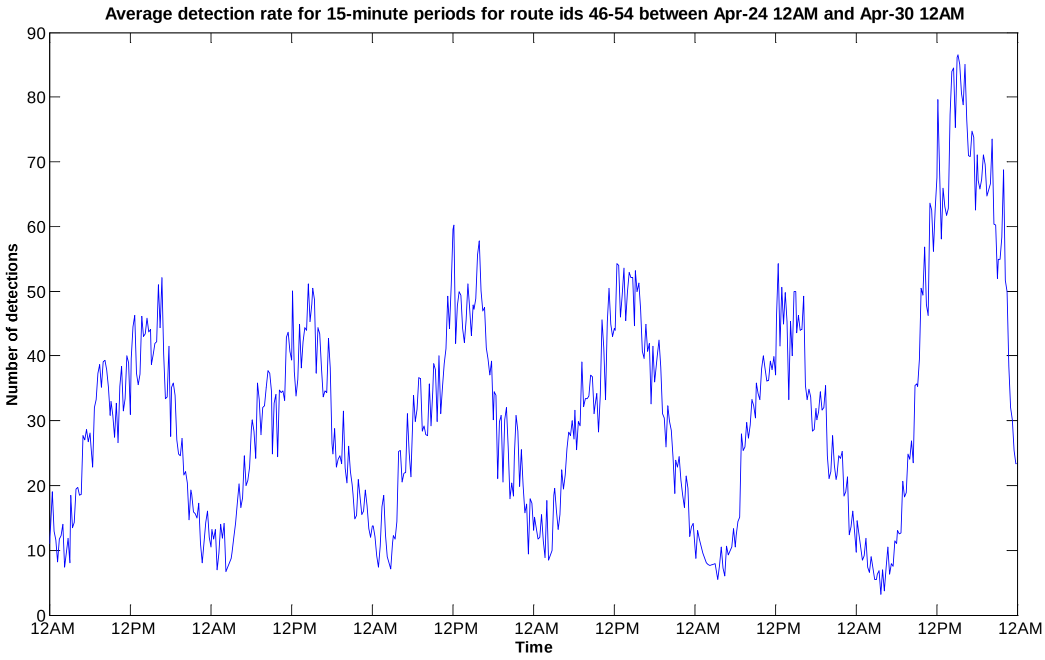


Figure 3-41: Bluetooth detection rate for highway I-15 Victorville southbound, week two

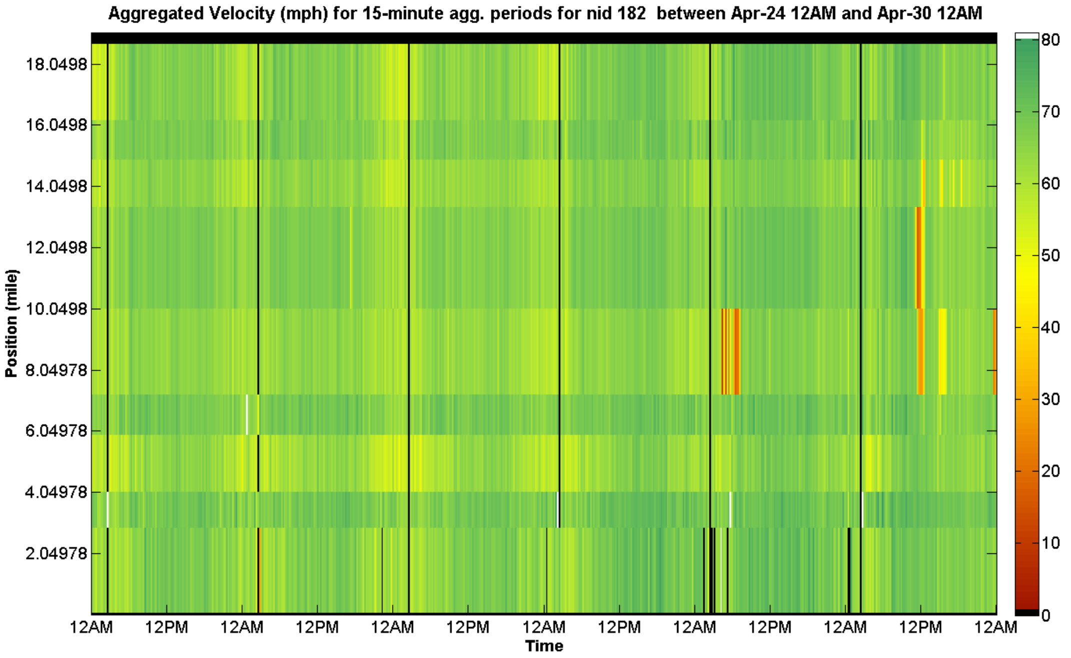


Figure 3-42: Time-space velocity grid for highway I-15 Victorville southbound, week two

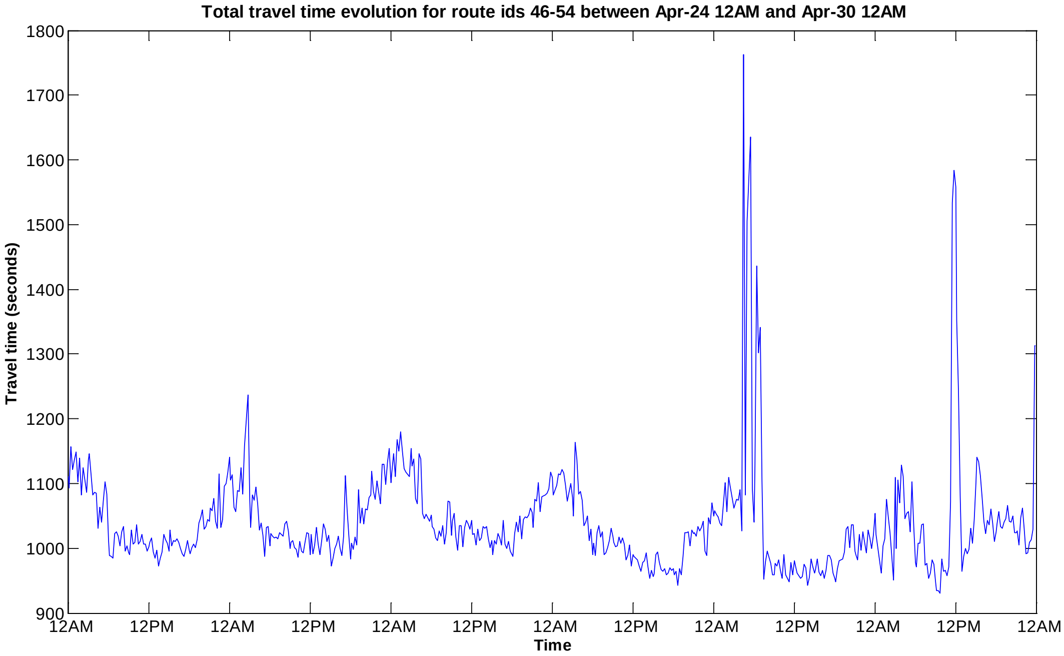


Figure 3-43: Total travel times for highway I-15 Victorville southbound, week two

2 BLUETOOTH SENSOR LOCATIONS

2.1 HIGHWAY I-15 ONTARIO

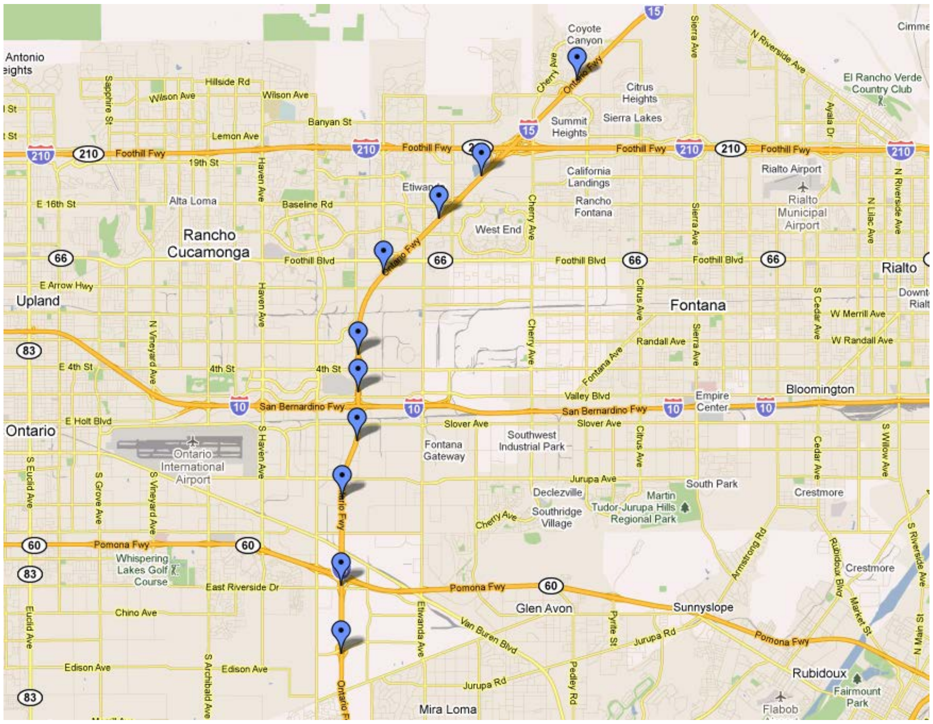


Figure 3-44: Map of Bluetooth sensors on highway I-15, Ontario

	1	2	3	4	5	6	7	8	9	10
1		2,5	3,6	5,0	6,6	7,3	8,2	9,3	10,9	12,2
2			1,1	2,5	4,1	4,8	5,7	6,8	8,4	9,7
3				1,4	3,0	3,7	4,6	5,7	7,3	8,6
4					1,6	2,3	3,2	4,3	5,9	7,2
5						0,7	1,6	2,7	4,3	5,6
6							0,9	2,0	3,6	4,9
7								1,1	2,7	4,0
8									1,6	2,9
9										1,26
10										

Figure 3-45: Approximate distances between sensors

BT1

Figure 3-46: Ontario Freeway, first light pole after Summit Avenue exit sign at Lat/Lng 34.154224, -117.474847

BT2

Figure 3-47: Ontario Freeway, penultimate light pole before “BaseLine Next Exit” sign at Lat/Lng 34.128846, -117.505590

BT3

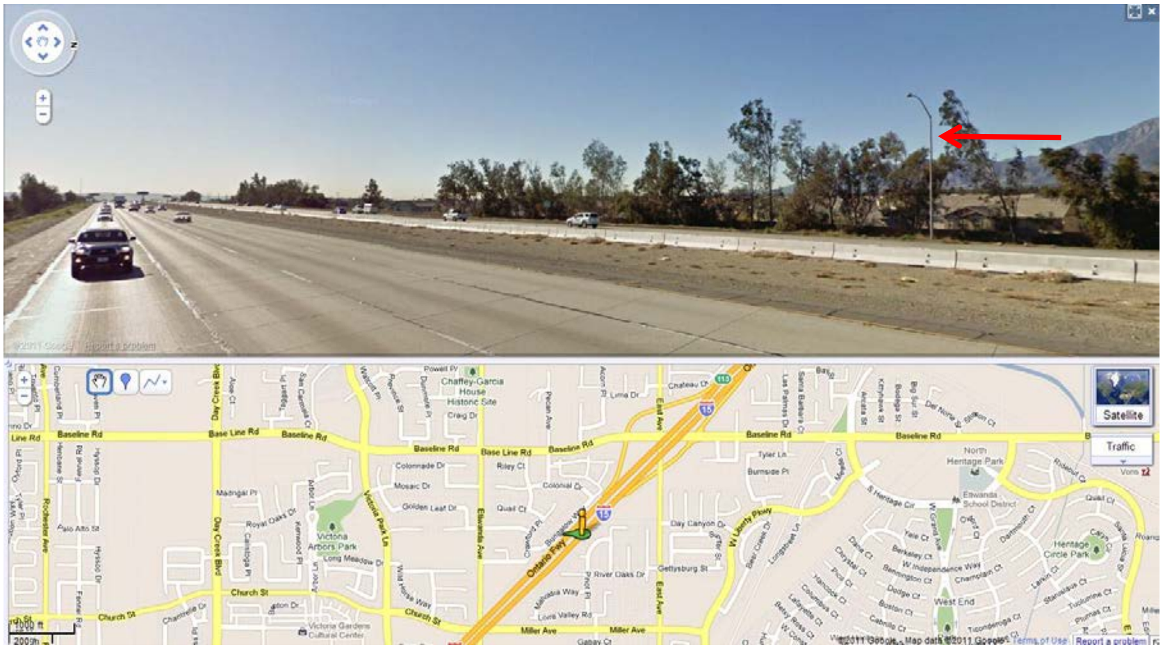


Figure 3-48: Ontario Freeway at Lat/Lng 34.117684, -117.518984

BT4

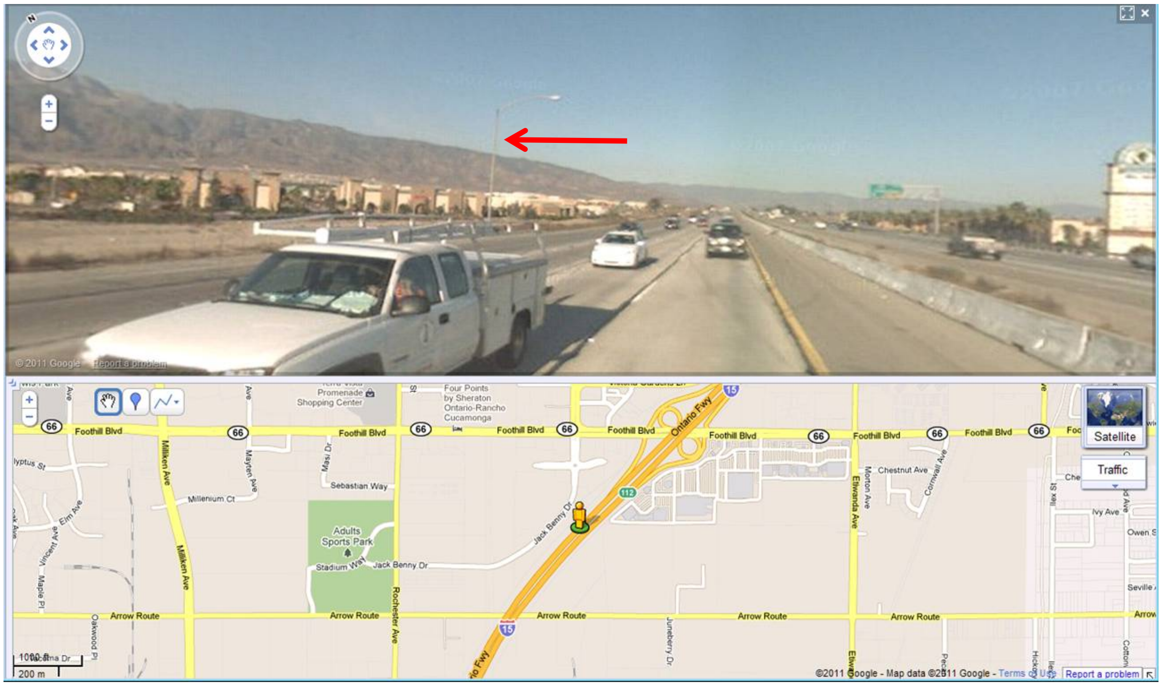


Figure 3-49: At Lat/Lng : 34.102928, -117.536678

BT5

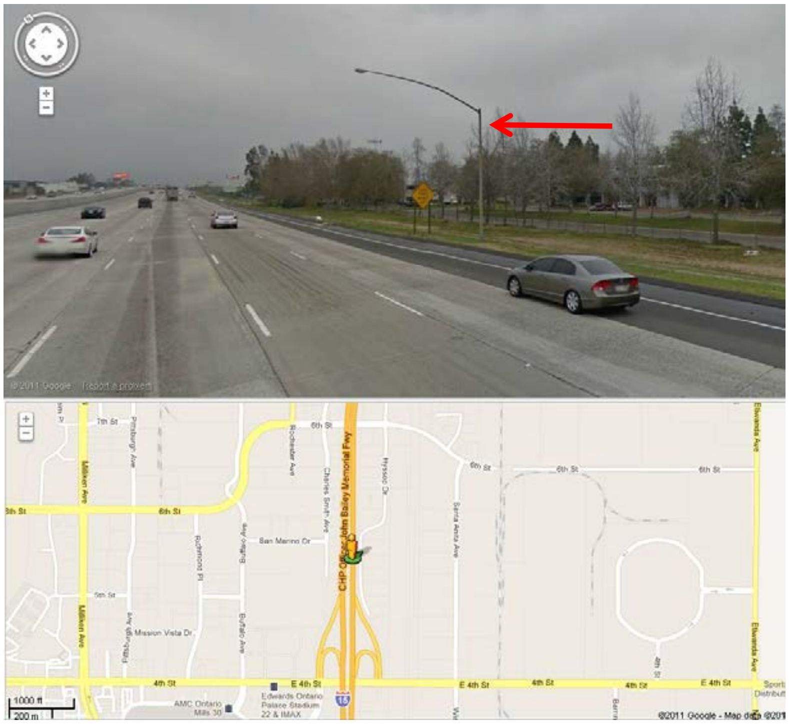


Figure 3-50: Last pole on northbound side after 4th street entrance; Lat/Lng 34.082544, -117.544540

BT6

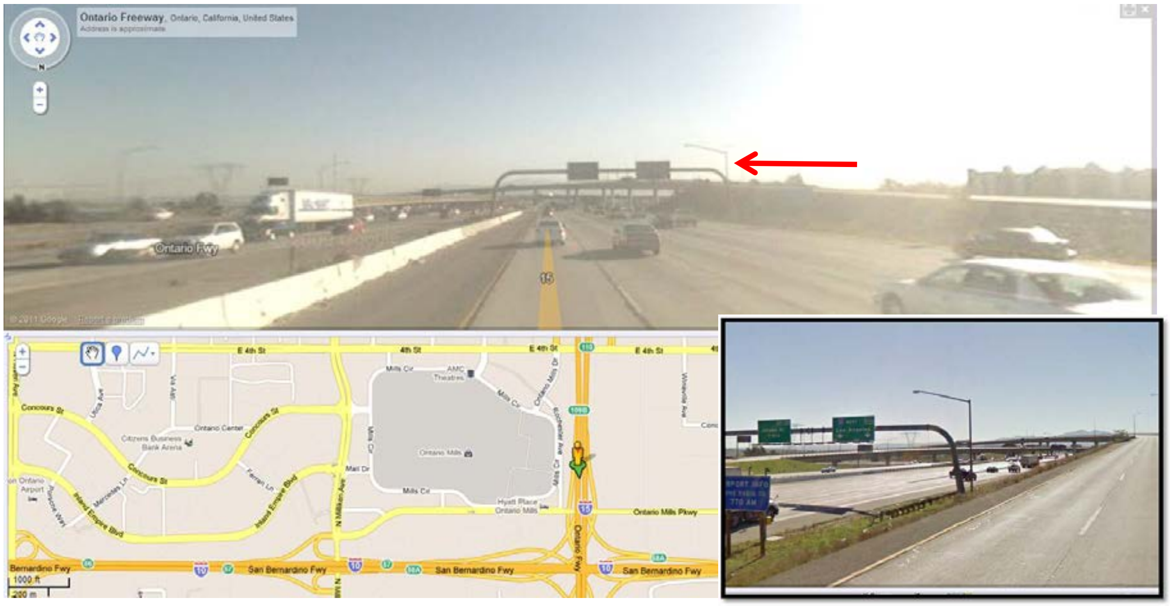


Figure 3-51: Ontario Freeway, next to the 108/109 exit sign at Lat/Lng 34.071191, -117.545297

BT7

Figure 3-52: Ontario Freeway, light pole just after Jurupa Street exit sign at Lat/Lng 34.058128, -117.545896

BT8

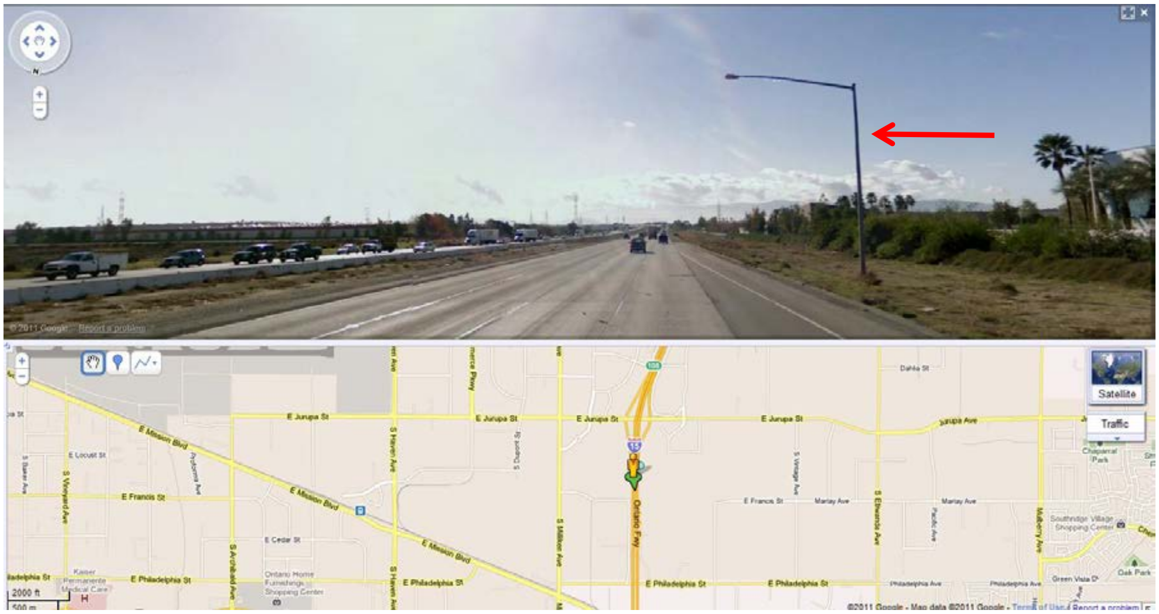


Figure 3-53: At the level of COVIDIEN building (on the right), at Lat/Lng 34.042582, -117.550531

BT9

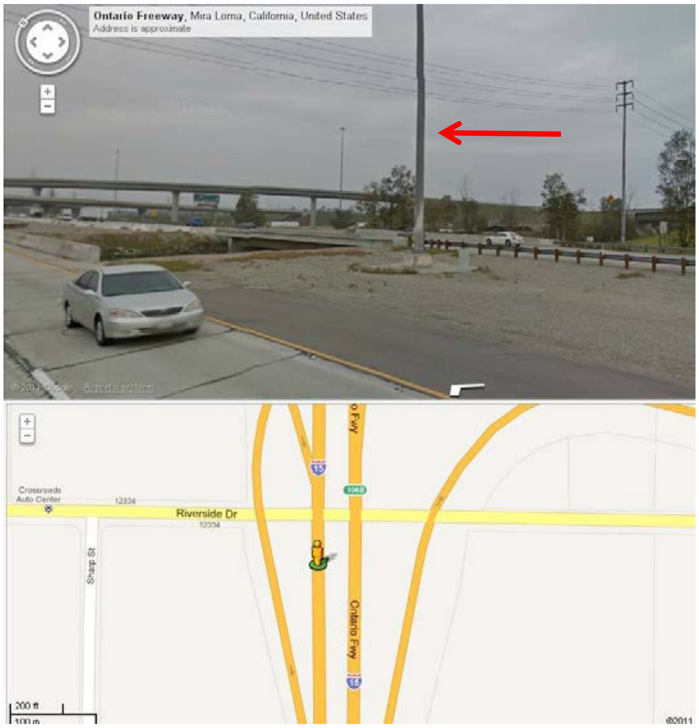


Figure 3-54: BT9 on huge pole just south of Riverside drive in the median of I-15

Lat/Lng 34.01851, -117.55031

BT10

Figure 3-55: Ontario Freeway, first light pole after Cantu-Galleano Ranch Rd overpass at Lat/Lng 34.000712, -117.550655

2.2 HIGHWAY I-15 VICTORVILLE

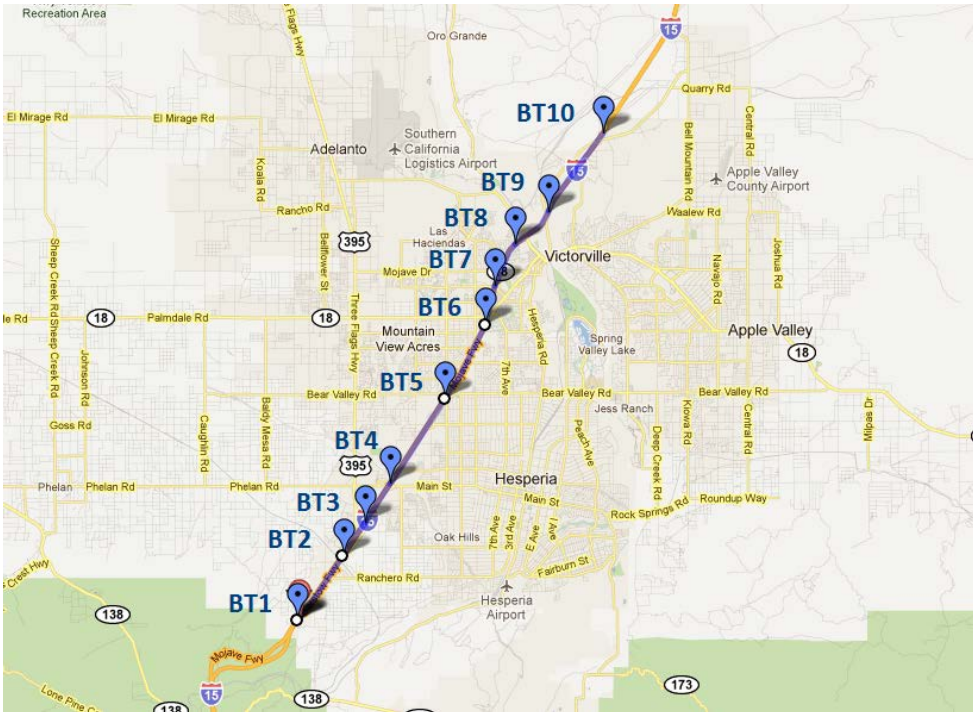


Figure 3-56: Map of Bluetooth sensors on highway I-15, Victorville

	1	2	3	4	5	6	7	8	9	10
1		2.61	3.9	5.44	8.75	11.48	12.83	14.36	15.91	19.04
2			1.29	2.83	6.14	8.87	10.22	11.75	13.3	16.43
3				1.54	4.85	7.58	8.93	10.46	12.01	15.14
4					3.31	6.04	7.39	8.92	10.47	13.6
5						2.73	4.08	5.61	7.16	10.29
6							1.35	2.88	4.43	7.56
7								1.53	3.08	6.21
8									1.55	4.68
9										3.13
10										

Figure 3-57: Approximate distances between sensors

BT1

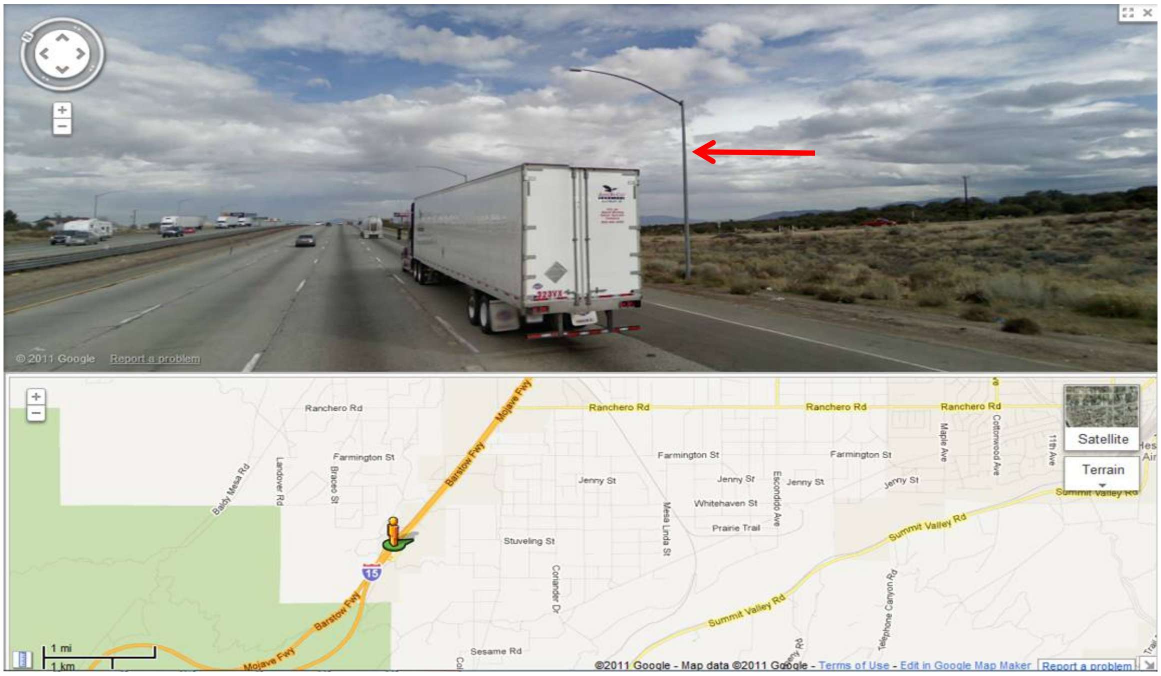


Figure 3-58: First light pole north to Oak Hill Rd overpass, at Lat/Lng 34.362674, -117.432518

BT2

Figure 3-59: Exit 141 to 395, Three Flags Highway at Lat/Lng 34.39328, -117.40664

BT3

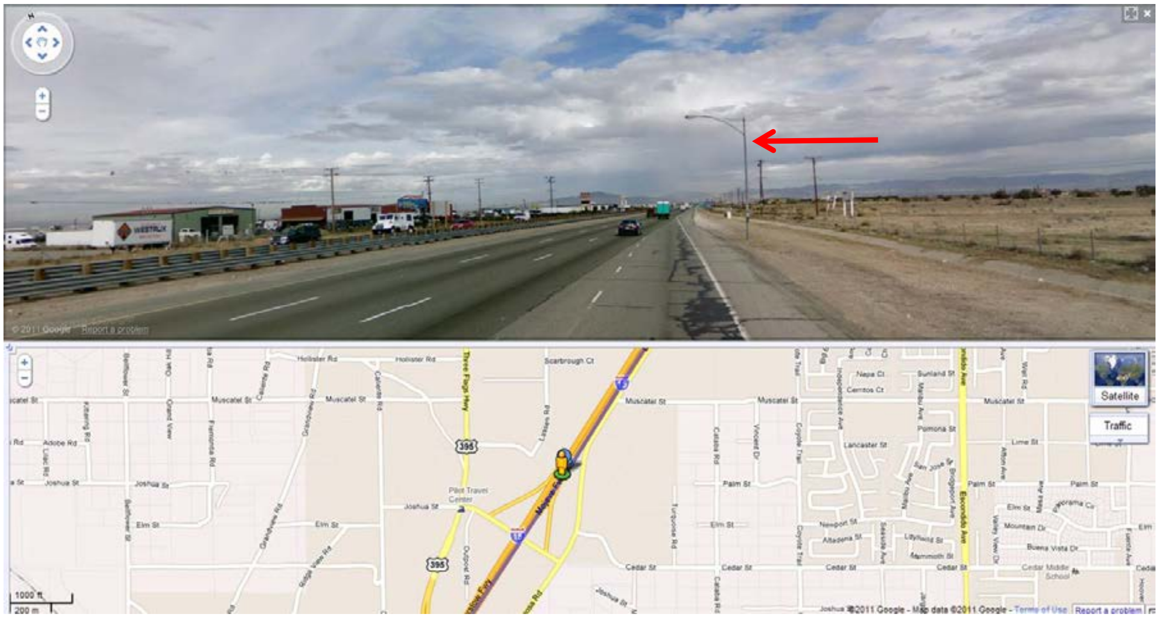


Figure 3-60: Last light pole on freeway on-ramp from Joshua Street at Lat/Lng 34.409253, -117.393816

BT4

Figure 3-61: Second pole after passing Main Street/Phelan Road overpass, at Lat/Lng 34.42789, -117.37925

BT5

Figure 3-62: Off-ramp to Mariposa, first pole before the other pole at level of exit 147 sign Lat/Lng 34.46834, -117.34767

BT6

Figure 3-63: Off-ramp to Palmdale Road, first light pole after Exit 18, at Lat/Lng 34.50310, -117.32418

BT7

Figure 3-64: Last pole before “Mojave D Exit Only” sign at Lat/Lng 34.52242, -117.31878

BT8

Figure 3-65: Exit 153A, at Lat/Lng 34.546305, -117.298772

BT9

Figure 3-66: At the level of Taco Chon Restaurant and Motel 6, at Lat/Lng 34.55678, -117.28803

BT10

Figure 3-67: At Lat/Lng 34.595012, -117.255980

2.3 HIGHWAY I-880

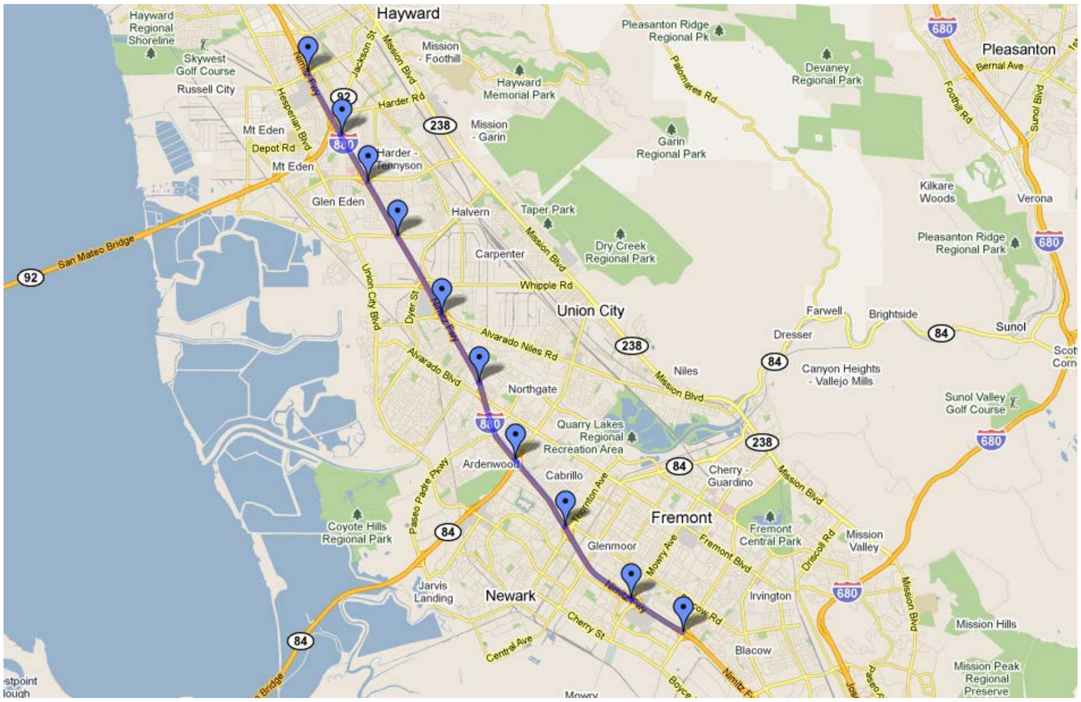


Figure 3-68: Map of Bluetooth sensors on highway I-880

	1	2	3	4	5	6	7	8	9	10
1		1,03	2,67	4,07	5,51	6,8	8,29	9,29	10,21	11,37
2			1,64	3,04	4,48	5,77	7,26	8,26	9,18	10,34
3				1,4	2,84	4,13	5,62	6,62	7,54	8,7
4					1,44	2,73	4,22	5,22	6,14	7,3
5						1,29	2,78	3,78	4,7	5,86
6							1,49	2,49	3,41	4,57
7								1	1,92	3,08
8									0,92	2,08
9										1,16
10										

Figure 3-69: Approximate distances between sensors

BT1

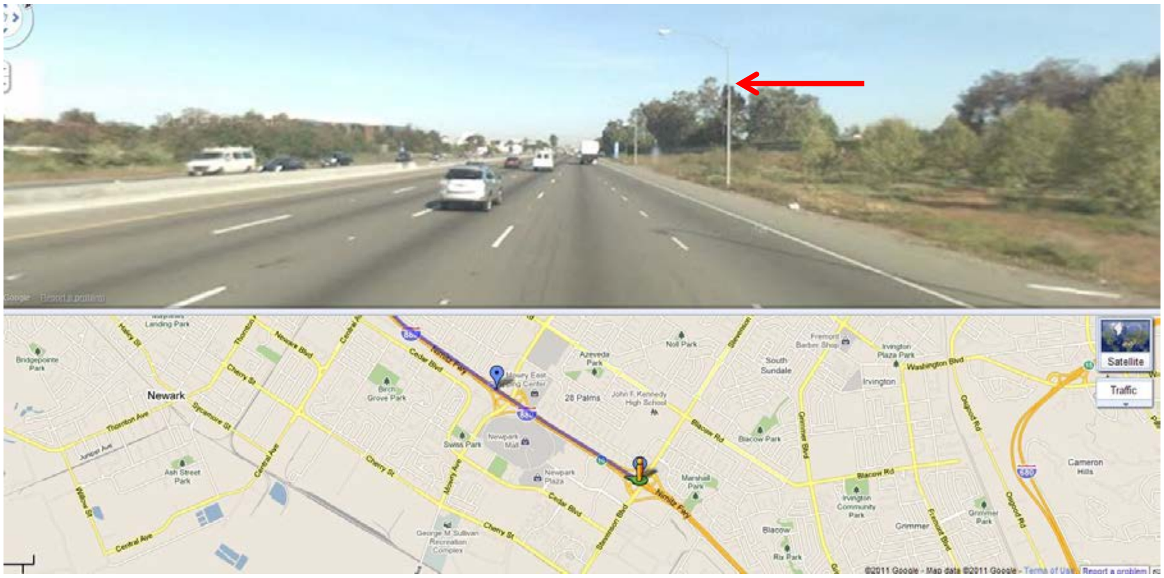


Figure 3-70: At Lat/Lng 37.523071, -121.988280

BT2

Figure 3-71: At Lat/Lng 37.530833, -122.003617

BT3

Figure 3-72: At Lat/Lng 37.548141, -122.023634

BT4

Figure 3-73: At Lat/Lng 37.564299, -122.038917

BT5

Figure 3-74: At Lat/Lng 37.5779, -122.0477

BT6

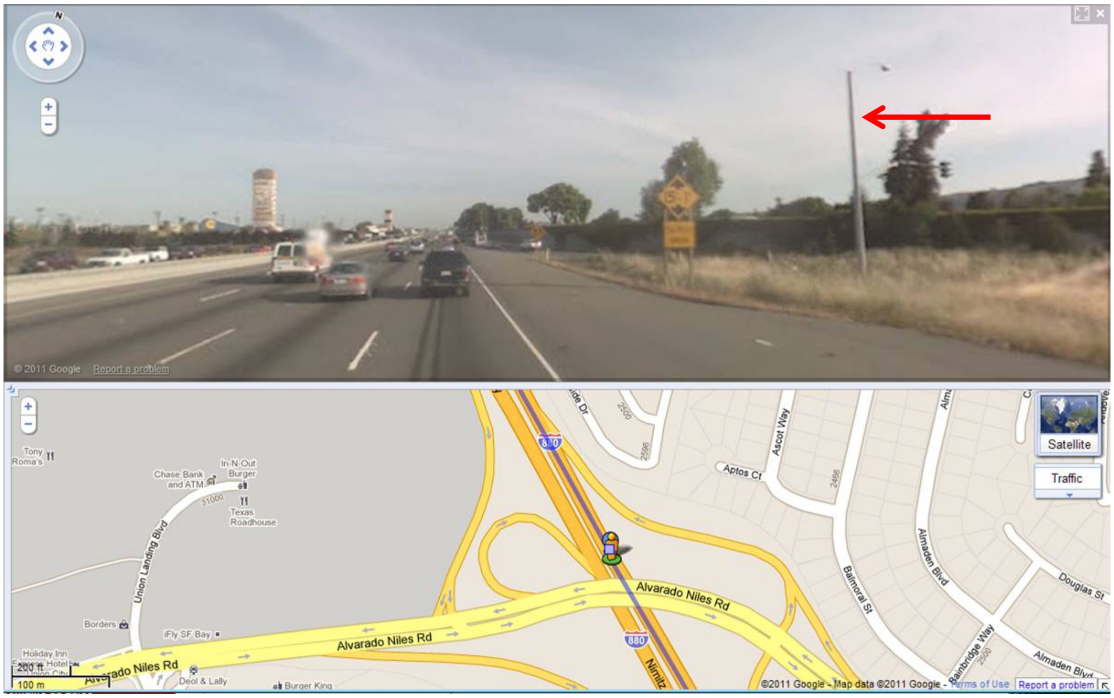


Figure 3-75: At Lat/Lng 37.599388, -122.060948

BT7

Figure 3-76: At Lat/Lng 37.618016, -122.074293

BT8

Figure 3-77: At Lat/Lng 37.630819, -122.083356

BT9

Figure 3-78: At Lat/Lng 37.6429, -122.0928, looking south. Pole is on southbound side of I-880 at interchange with CA 92 (built recently), south of the overpasses.

BT10

Figure 3-79: At Lat/Lng 37.656909, -122.101335

3 THE BLUETOOTH STATION FILTER: ANALYSIS OF THE PARAMETERS

One of the challenges in determining travel times with Bluetooth sensors is the problem of multiple detections, where a vehicle is detected more than once by the same sensor as it passes through that sensor's detection field. As described in section 1.5.1, the Bluetooth station filter is an algorithm that addresses that problem by eliminating duplicate entries for the same vehicle (within a single trip) at a given sensor.

Basic filtering strategies include:

- **First time**: Using the first point in time a vehicle is detected by a given sensor works best for multiple readings due to off-ramps. This is important because an off-ramp can slow down traffic in the rightmost lane of the highway, and what we're really interested in is the travel-time for mainline flow along the highway.
- **Last time**: Using the last point in time a vehicle is detected by a given sensor works best for multiple readings due to on-ramps. By picking the last time, we eliminate the time spent on the ramp and get a more accurate estimate. This is especially important at metered lights.
- **Average time**: Using the average time of the measurements should give us an estimate of the time at which that vehicle was closest to the sensor if the multiple readings were simply due to congestion. This would be the strategy of choice in a case where there are no nearby ramps.

Ultimately, choosing the average time turned out to be the best strategy, and this is what was done for results shown in all other parts of this report. In this section, we present the methodology we used for choosing the best strategy for filtering the multiple detections at each of the Bluetooth stations. This approach relies on uniformity measures which are themselves dependent on the standard deviations of the travel times over some discretization of the time and space. These measures, defined in this document, are the basis of an analysis enabling us to state that a given strategy is better than another in some cases.

As part of the methodology, we define and then compare three strategies: the "first time" detection strategy, the "average time" detection strategy, and a strategy that we called "adapted" because we chose the strategy for each Bluetooth sensor based on their locations and their road environments (on-ramps, metered lights, off-ramps). The methodology allows us to strictly compare those strategies in order to keep the best in the implementation of the whole filter.

3.1 METHODOLOGY

3.1.1 OBSERVATIONS

When we originally ran the entire Bluetooth filter on the three Bluetooth deployments, little was known about the best strategy choice for the Bluetooth station filter and its multiple detection problem. When there was a multiple detection of the same MAC id, the “first time” detection was arbitrarily chosen to be the detection time of this MAC id at this Bluetooth sensor. However, after looking at the travel times over time and space after all the filtering and aggregations had been done, a flaw was detected in this choice of data processing. Indeed, we can see in Figure 3-80 a constant mean speed difference between consecutive routes (horizontal color bands). The colors on the map indicate miles/hour, corresponding to the scale on the right side of the figure.

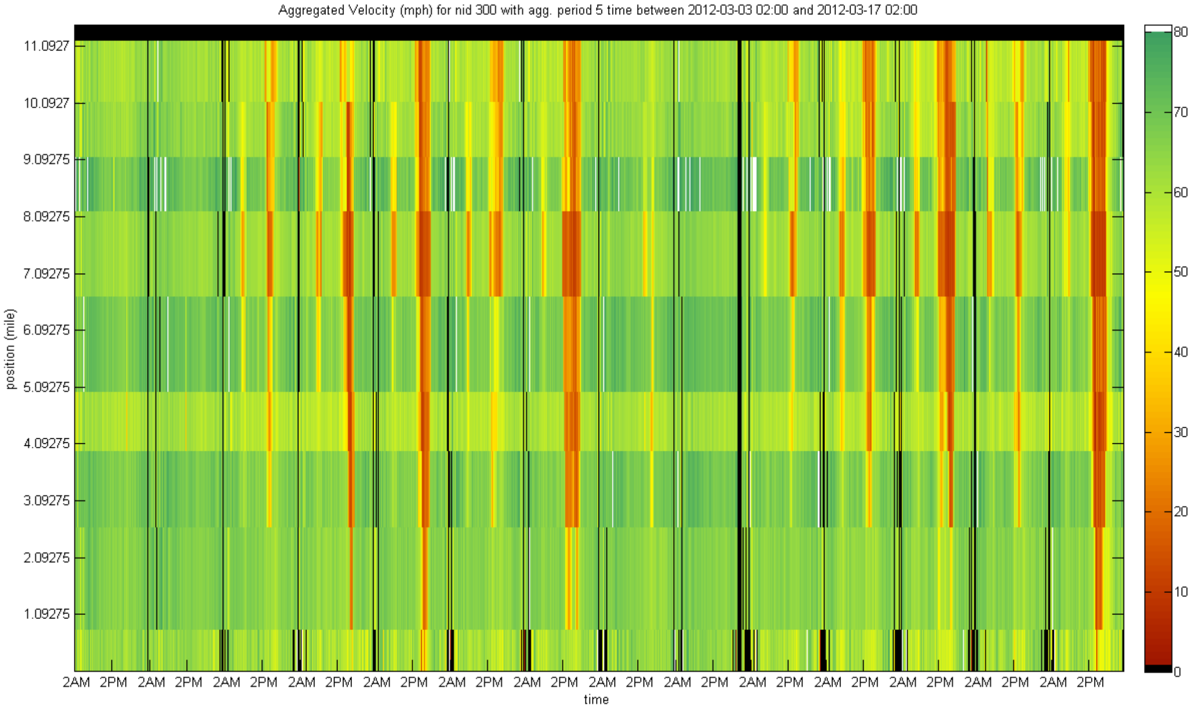


Figure 3-80: Bluetooth aggregated velocity output on I-880 northbound for the “first time” strategy

This results in distinct color bands that are hard to explain. In fact, in free-flow we are not expecting the mean speeds to be that different between consecutive routes. The Bluetooth output seems to imply an accordion behavior from the average driver, constantly decelerating and accelerating. It seems unlikely that the physics of the road make the drivers drive like that. If we number the routes from 1 to 9 as the position increases, a quick analysis would result in saying that the data processing is obviously overestimating the speeds and underestimating travel times in routes 3, 5, and 7 while underestimating the speeds and overestimating travel times in routes 1, 4, 6, 8, and 9.

3.1.2 MULTIPLE DETECTION STRATEGIES AS PARAMETERS OF THE FILTER

In re-examining the implementation of our Bluetooth filter, we decided to assign one strategy to each of the distinct associations of Bluetooth sensors and directions, that we call stations. From now on we will use the term “time tag” because the name strategy will be used to designate the global choice of all the chosen time tags for all the Bluetooth stations of a specific experiment. Theoretically, a sensor time tag could be time-evolving, but we do not consider that in the context of this document. Further research would be needed in order to choose the appropriate time tag as a function of time and state of the road.

On I-880 ten Bluetooth sensors were deployed during two weeks of March 2012 (from March 2 to March 17). If we were studying both directions that would mean 20 stations for which we want to choose the best time tags. Considering only the northbound directions, there are only 10 stations to look at, and a possible strategy is to choose the “first time” time tag for each of the stations. Another option could be to assign the “average time” time tag to each of the sensors, or to assign the “first time” time tag to the five first sensors and the “last time” for the remaining ones. The next section defines measures of strategy quality.

3.1.3 MEASURING STRATEGY QUALITY

We needed to develop a way of comparing the quality of these strategies and do so for each of our Bluetooth deployments. We devised the following definitions:

- **Definition 1.** A Bluetooth **route** is the part of the highway between two Bluetooth sensors. With ten sensors on a section of highway we have 18 routes, nine going northbound and nine going southbound.

- **Definition 2.** We say that strategy 1 is **better** than strategy 2 **for a particular route** if the standard deviation of the travel times of strategy 1 is consistently smaller than that of strategy 2 over time. In other words, strategy 1 is better for a particular route if its travel time distribution is more uniform.

- **Definition 3.** We say that strategy 1 is **globally better** than strategy 2 if the aggregated paces (or speeds) of strategy 1 over the routes are more uniform than those of strategy 2. The uniformity of these travel times is another time assessed by first computing the variance of these aggregated paces for each five minutes per mile (defined by the fixed aggregation period) and then averaging these variances over time. A smaller mean variance means more uniformity.

- **Definition 4.** We say that strategy 1 is **better** than strategy 2 if strategy 1 is **better** than strategy 2 for a large majority of routes **AND** also **globally better**.

3.2 RESULTS

The total number of strategies is 3^{10} for each Bluetooth deployment. Due to time constraints of this project we were able to focus only on the I-880 Bluetooth deployment and the three following three strategies:

- **“First time” strategy**—The “first time” time tag is applied to all the Bluetooth sensors.
- **“Average time” strategy**—The “average time” time tag is applied to all the sensors.
- **“Adapted” strategy**—This is an attempt to devise a strategy which is more adapted to the physics of the road, and more particularly to the locations of on-ramps and off-ramps.

Each of the concepts of strategy quality defined above relies on metrics aggregated over time and space. The routes are already fixed by the locations of the sensors, and we chose to set the aggregation period to five minutes.

3.2.1 “FIRST TIME” VERSUS “AVERAGE TIME”

Comparison for each route

Let us call strategy 1 the “First time” strategy and strategy 2 the “Average time” strategy. Table 2 shows the average difference of standard deviations between strategy 1 and 2 for each route of I-880 northbound.

Table 2: Average standard deviation difference between strategy 2 and 1

Route number	(Average std strategy 2) – (Average std strategy 1)
1	-1.6057
2	-0.5522
3	-0.5413
4	-1.2691
5	-0.5014
6	-0.0984
7	-0.9266
8	-0.9317
9	-0.6242

Instead of computing only the average, it is also possible to look more precisely and plot the evolution of the difference of standard deviations over time for each route number. Figure 3-81 shows the result for route number 1. The figures for all the routes are available in section 3.4. All these figures indicate a smaller standard deviation for strategy 2 for almost every time period and for each route (except route 6 where the standard deviations are comparable). We can then conclude that **strategy 2 is better than strategy 1 for each route** except for route 6 where the two strategies are equivalent.

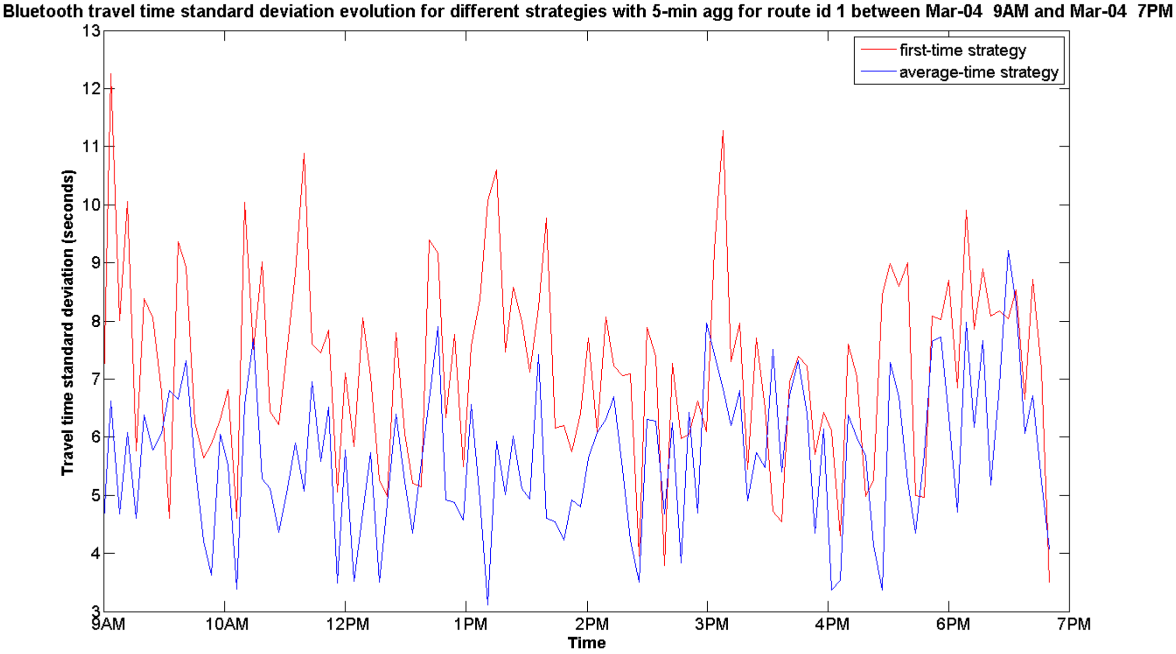


Figure 3-81: Standard deviation evolution for different strategies

Global uniformity over routes

Here we define the concept of global uniformity and one strategy being globally better than another. The measure used for global uniformity is computed in the following way:

- For each route, the travel times are aggregated according to a regular discretization of the time which is defined by an aggregation period. After dividing by the route length, this gives for each five minutes per mile an average speed over the route.
- For each five minutes per mile, we compute the variance of this list of average speeds that we call global standard deviation for this strategy and this five minutes per mile.
- This variance is then plotted over time (see Figure 3-82) and is also averaged to give a number which represents the global uniformity of this strategy.

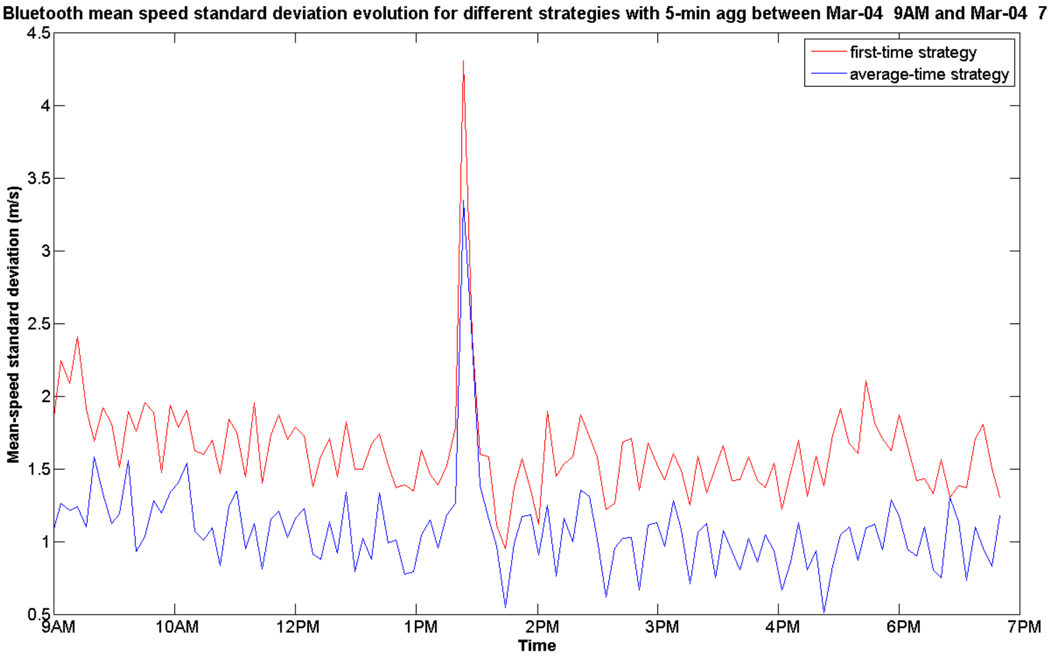


Figure 3-82: Global standard deviation evolution for different strategies

We can say from Figure 3-82 that **strategy 2 is globally better than strategy 1**. Because strategy 2 is also better than strategy 1 for each particular route, we can then conclude that **strategy 2 is better than strategy 1.**

An attempt to adapt a priori the strategy to the route

At the very beginning of creating the Bluetooth filter, the pros and cons of each station strategy was already described as a non-trivial problem:

- **Average time:** Picking the average time of the measurements should give an estimate of the time at which that vehicle was closest to the sensor if the multiple readings were simply due to congestion. This would be the strategy of choice in a case where there are no nearby ramps.

- **Last time:** Picking the last time of measurement works best for multiple readings due to on-ramps. By picking the last time we eliminate the time spent on the ramp and get a more accurate estimate. This is especially important at metered lights.

- **First time:** Picking the first time of measurement works best for multiple readings due to off-ramps. This is important because an off-ramp can slow down traffic in the rightmost lane of the highway, and what we're really interested in is the travel-time for mainline flow along the highway.

After implementing the Bluetooth filter and looking at the results, the problem is more mature and we can add some notes to this early description. First, because we are mainly interested in congestion but at the same time want to be precise during free-flow, taking the average time should be the default choice, so it was actually wrong to take the first-time detection strategy as default in the first place. Second, looking into more details at the actual positions of on-ramps (eventually with metered lights) and off-ramps could help us in adapting the strategy for each station. Third, it seems that the seemingly overestimated travel times in some of the routes could be the result of a nearby on-ramp or off-ramp. The seemingly underestimated travel times in some of the routes could be simply because these routes are just after routes that have overestimated travel times. The travel times could be indeed underestimated or just seem to be in contrast with the overestimated travel times. Finally, the outlier filter should filter the obvious outliers, but some measurements coming from on-ramps or finishing at off-ramps may not be considered outliers.

With Google Maps and Google Street View we looked more closely into the road environment of the Bluetooth sensors in the northbound directions. The resulting screenshots can be seen in section 3.5. In Table 3, we say that there is an on-ramp influencing a Bluetooth station if this on-ramp is going northbound and located within the 100-meter range of the Bluetooth sensor.

Table 3: Map environment of Bluetooth northbound stations. “Under” stands for visually underestimated speed, “over” for overestimated speed.

Station	On-ramp	Metered light	Off-ramp	End of route	Start of route
1	yes	yes	/	/	1, under
2	yes	yes	yes	1, under	2
3	yes	yes	no	2	3, over
4	yes (2)	yes	yes (1)	3, over	4, under
5	yes	/	yes	4, under	5, over
6	yes (on it)	yes	no	5, over	6, under
7	yes (really close)		no	6, under	7
8	yes		yes (on it)	7, over	8, under
9	no		yes	8, under	9, under
10	yes		yes	9, under	/

Following the analysis performed earlier in this section, we would advise the following strategy:

1. **Last time for detector 1** because there is an on-ramp AND a metered light. That should underestimate the travel time for route 1 a bit less (meaning the green for this route will be darker).
2. **Last time for detector 2** because it is closer to the on-ramp than it is to the off-ramp AND there is a metered light.

3. **Last time for detector 3** because it also is closer to the on-ramp than it is to the off-ramp AND there is a metered light.
4. **Last time for detector 4** because it is close to two on-ramps AND there is a metered light. The distance to the off-ramp is equivalent to the distance to the two on-ramps.
5. **First time for detector 5** because it is closer to the off-ramp than it is to the on-ramp, and it is not close to a metered light. That should make route 5 lighter, which is good because currently it is visually overestimating speed.
6. **Last time for detector 6** because it is on an on-ramp, not close to the off-ramp, AND there is a metered light. (Google Street View even shows some congestion there.)
7. **Last time for detector 7** because it is on an on-ramp, there is no off-ramp, AND there is a metered light. This should make route 7 lighter.
8. **First time for detector 8** because it is closer to an off-ramp, but this would need more research/testing.
9. **First time for detector 9** because of the off-ramp.
10. **First time for detector 10** because of the off-ramp.

The next section discusses the same analysis but this time compares the “average time” strategy to the “adapted” strategy (more adapted to the physics of the road, and more particularly to the locations of on-ramps and off-ramps).

3.2.2 “AVERAGE TIME” VERSUS “ADAPTED”

Comparison for each route

Let us call strategy 2 the “Average time” strategy and strategy 3 the “Adapted” strategy. Table 4 shows the average difference of standard deviations between strategy 2 and 3 for each route of I-880 northbound.

Table 4: Average standard deviation difference between strategy 3 and 2

Route number	(Average std strategy 3) – (Average std strategy 2)
1	-1.4552
2	1.1987
3	-0.4182
4	-0.9452
5	0.3989
6	-0.2282
7	-0.7840
8	-0.5054
9	-0.6242

It seems that **strategy 2 is still better on average than strategy 3** for most of the routes, except for routes 2 and 5. This, then, is the first indicator of the failure of our attempt to adapt the time tags to the particularities of the road. Figure 3-83 and Figure 3-84 show the evolution of the travel time standard deviations for routes 1 and 2, respectively. The remainder of the figures (for routes 3 to 9) can be found in section 3.6.

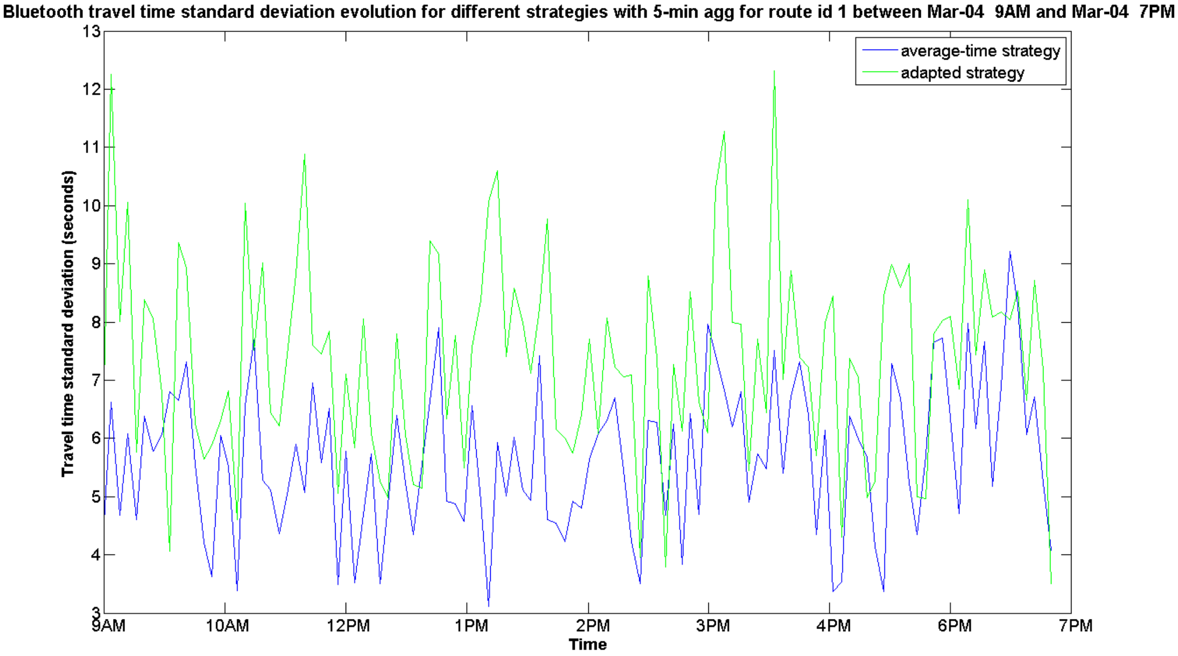


Figure 3-83: Standard deviation evolution for route 1, strategies 2 and 3

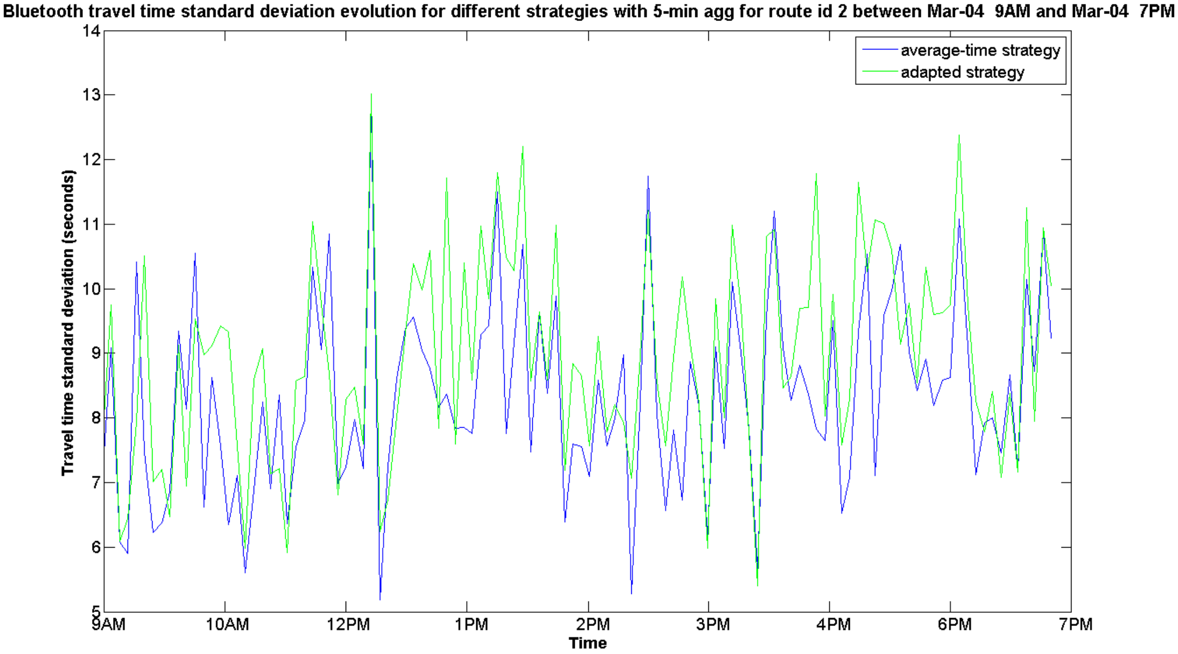


Figure 3-84: Standard deviation evolution for route 2, strategies 2 and 3

Global uniformity over routes

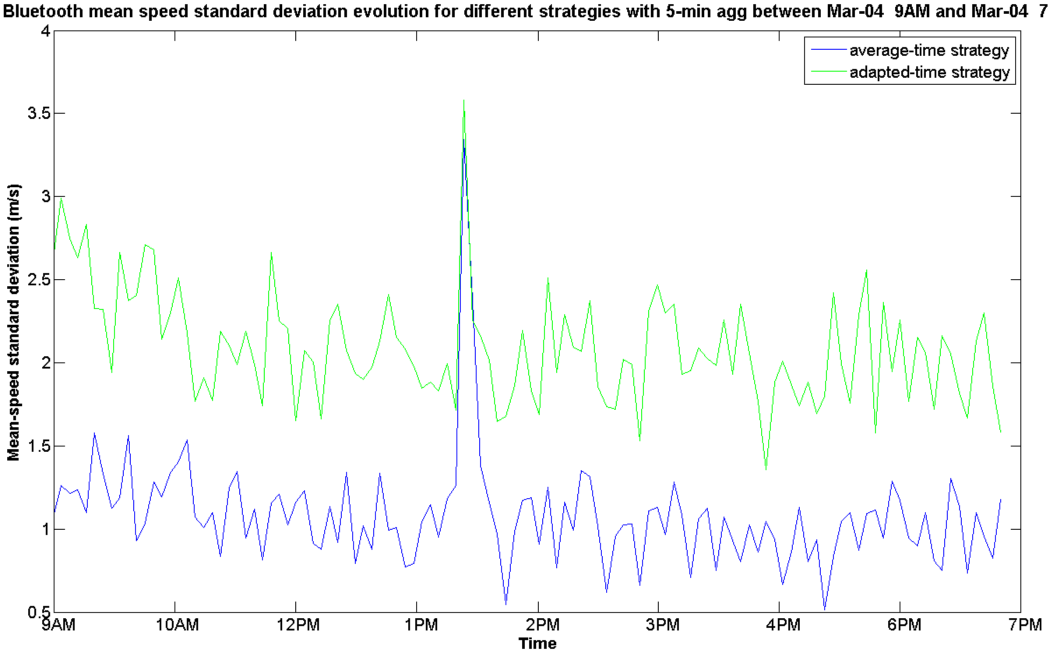


Figure 3-85: Global Standard deviation evolution for strategies 2 and 3

We can say from Figure 3-85 that strategy 2 is globally better than strategy 3.

Because strategy 2 is also better than strategy 3 for the majority of the routes, we can then conclude that **strategy 2 is better than strategy 3**. Therefore our new “Adapted” strategy ends up performing worse than the “Average time” strategy for our choice of uniformity measures. Many other strategies could be tested against the “Average time” strategy in order to get the best one. However, due to time constraints we decided to go forward with the average time strategy for the three Bluetooth deployments.

3.2.3 DIFFERENCES IN SPEED (OR PACE) OUTPUT

Our methodology for comparing strategy quality defines and computes different uniformity measures using only the travel time standard deviations. Even though they aren’t part of the uniformity measures, however, the aggregated travel times themselves are obviously influenced by the chosen strategy. This section shows how these travel times are modified. To make the analysis visually more intuitive, we convert travel times to velocities and plot the results as a velocity map, as in Figure 3-80 and Figure 3-86.

Classification of routes

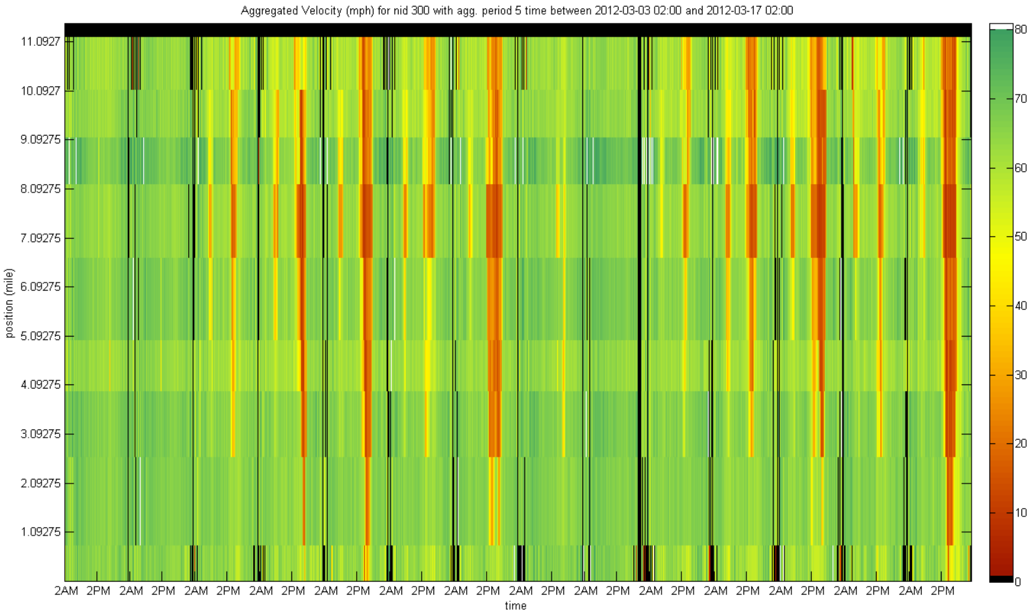


Figure 3-86: Bluetooth aggregated travel times (displayed here as velocities) for the "average time" strategy on I-880

Looking at Figure 3-86, it is possible to make a visual assumption that classifies the routes into "underestimated" and "overestimated" routes in terms of velocities. The results of this quick visual analysis are shown in Table 5.

Table 5: Visual assumption on route speed accuracy

Route number	Under/Over
1	Under
2	Under
3	Over
4	Under
5	Over
6	Under
7	Over
8	Under
9	Under

While we realize that this visual assessment does not provide any proof about which route is actually underestimated or overestimated, we would still feel more comfortable about our analysis if the result of our best strategy is a decrease in the mean speed for “overestimated” routes and an increase in the mean speed for “underestimated” routes.

In section 3.6 we plot the average speed evolution for the three strategies considered in this study and for each route. The aggregation period used for averaging the speeds over time is five minutes. It can be observed from these figures that for each route except route 7, the strategy that we consider the best as a result of our study (strategy 2, “average time”) is the one that decreases the most the average speeds for “underestimated” routes (1,2,4,6,8,9) and increases the most the average speeds for “overestimated” routes (3,5,7).

This is a good result because if confirmed with other simulations (other days, other sections of highway) it would mean that there is actually a correlation between the uniformity measures that we chose and the resulting average speeds, and that this correlation goes in the right direction, meaning that minimizing the uniformity measure will actually tend to re-equilibrate the routes’ travel times. Visually, this will make differences in color between the roads less pronounced in Figure 3-80, for example.

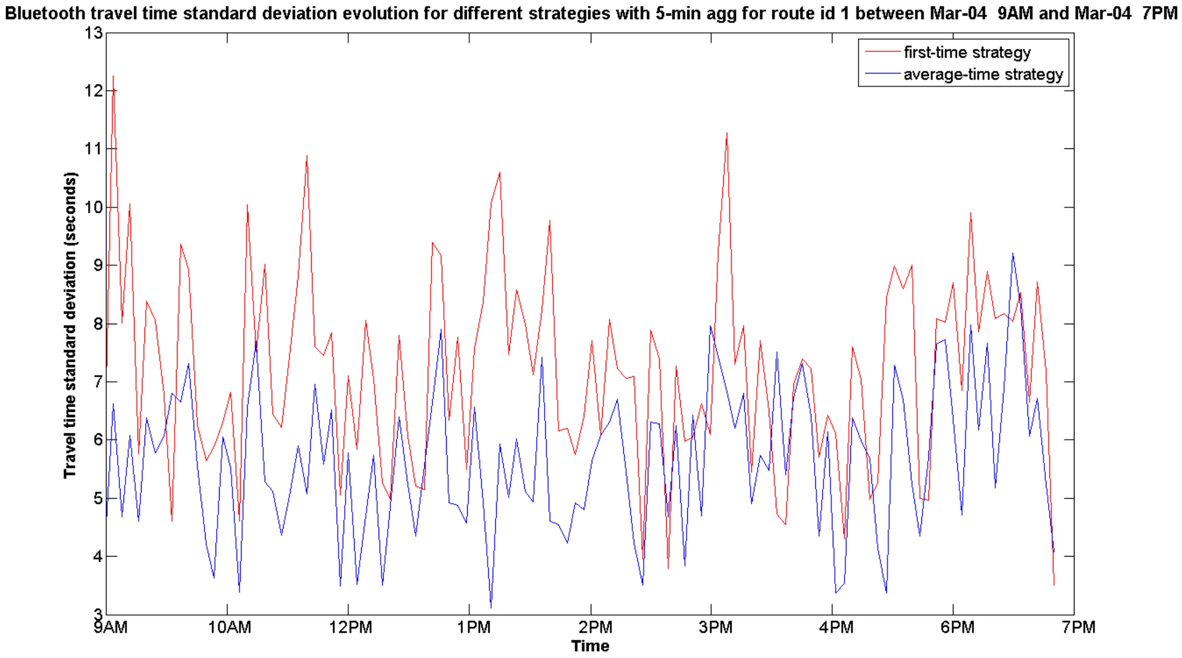
For route 7 where the mean speed is visually overestimated, the strategy that would decrease the speeds the most is the “Adapted time” strategy. However, this can be only due to the fact that for this strategy we chose to take the last-time detection at sensor 7 and the first-time detection at sensor 8. Further research would need to be done to see if we could infer a better strategy than “average time” from this observation (for example by choosing the last-time detection at sensor 7 or the first-time detection at sensor 8 or both).

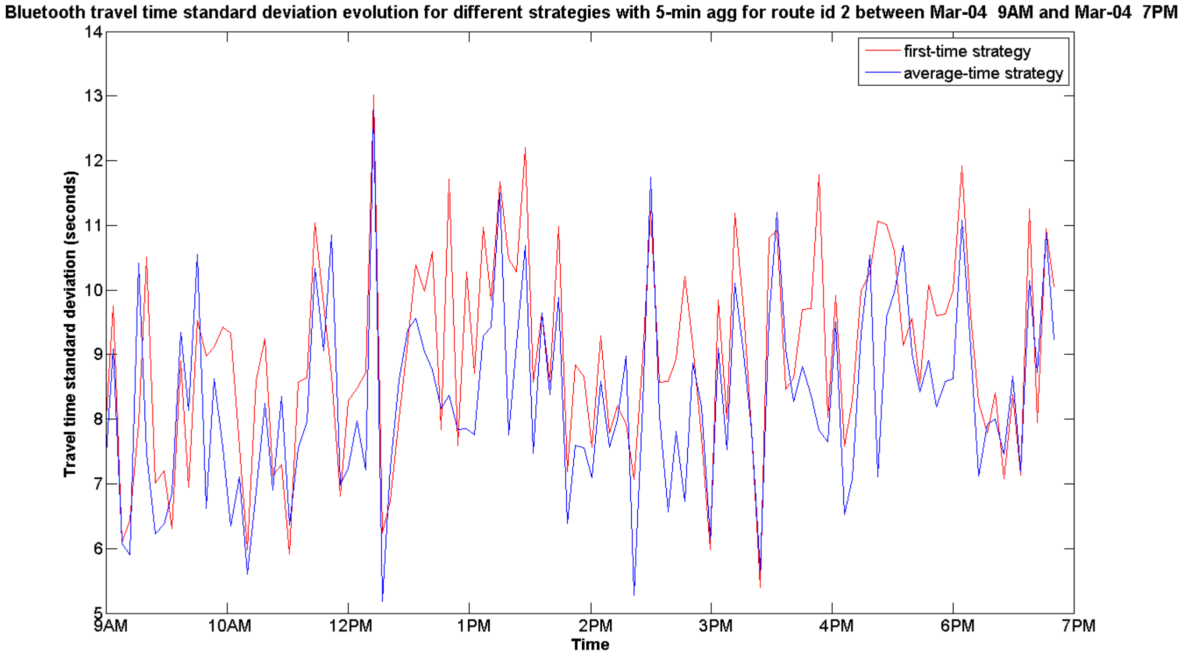
3.3 CONCLUSION

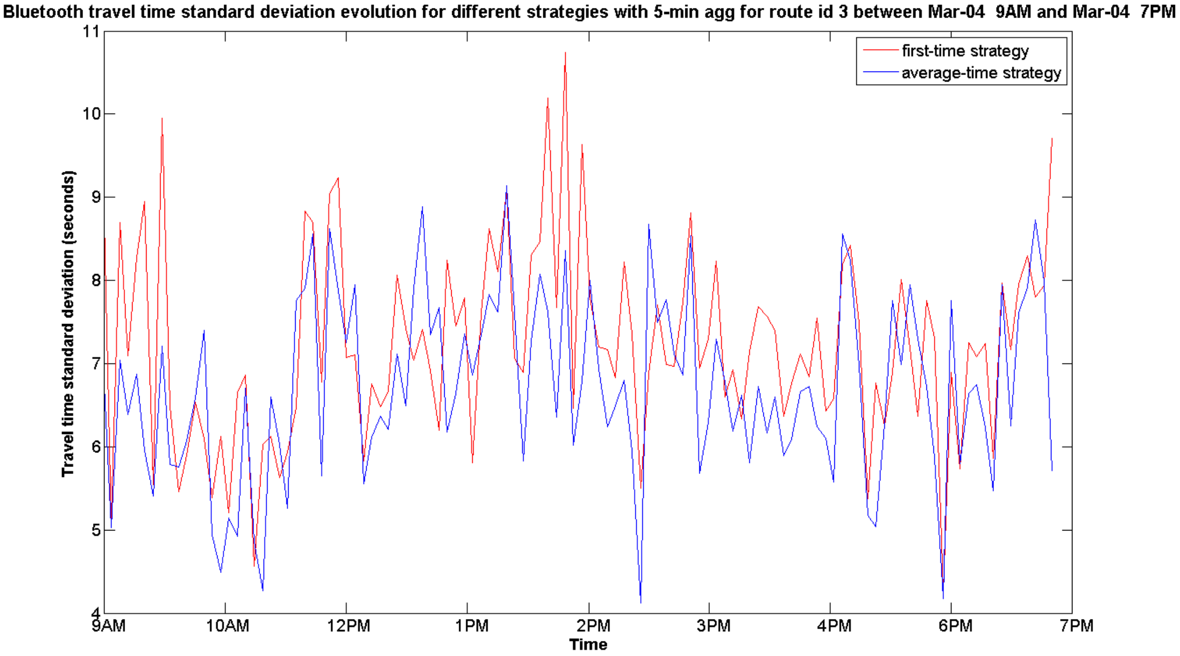
We started from the observation that there were unexpected differences between the free flow average speeds on different routes. We postulated that this was partly due to the choice of the parameters of the station filter, that is, what should be used as the observation time when the same Bluetooth device was detected multiple times by the same Bluetooth sensor. We defined the notion of strategy, which represents a certain choice of the parameters of the station filter. To be able to compare the strategies we defined two different measures of uniformity, each of these relying on the variance of aggregated travel times, and said that one strategy was better than another if its values for these two measures were smaller than those of the other strategy. This gave us a methodology for finding the best strategy to use with the Bluetooth station filter on a specific section of highway. We then applied this methodology to compare three strategies on a section of highway I-880 in the northbound direction. The best strategy was then chosen to be applied to the Bluetooth station filter in this direction. Further work would involve trying more strategies for this specific location, to try to find a strategy better than what we considered the best, and also to test this methodology on other times and places.

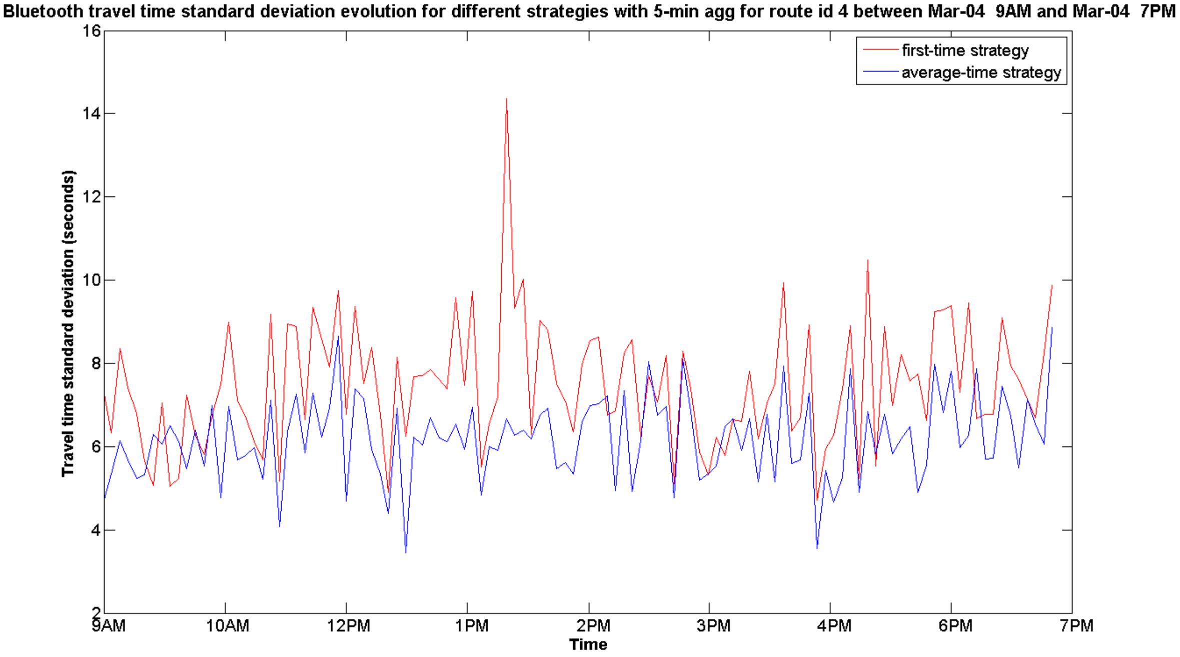
3.4 STANDARD DEVIATION EVOLUTIONS FOR DIFFERENT ROUTES AND DIFFERENT STRATEGIES

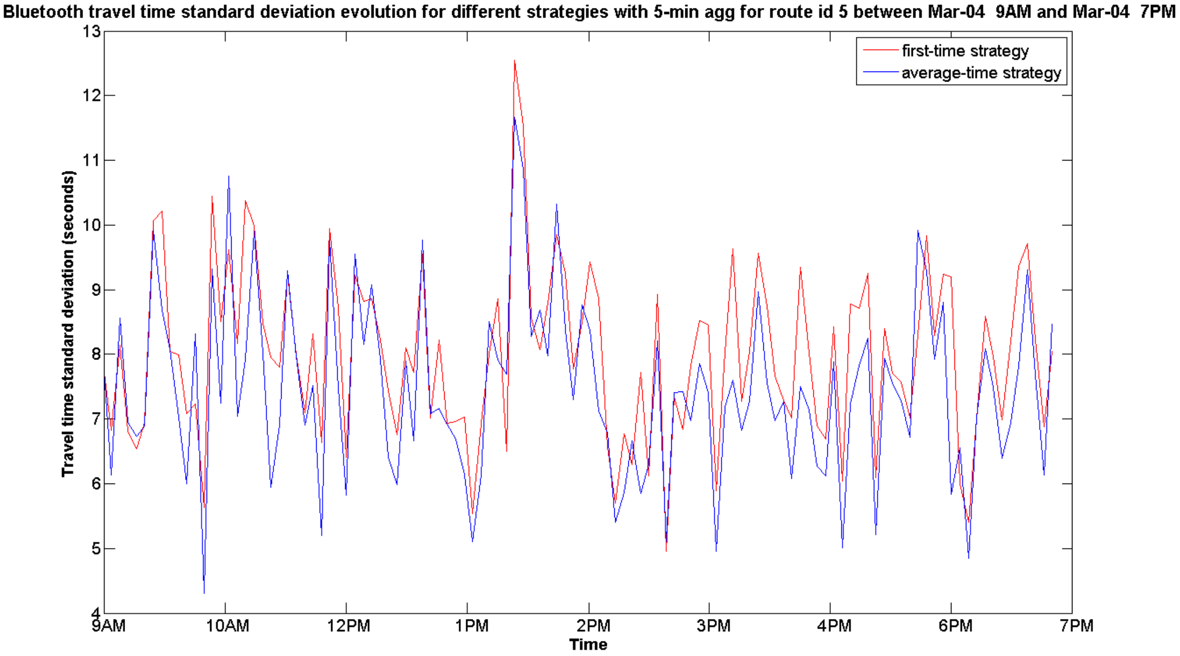
“First time” versus “Average time”

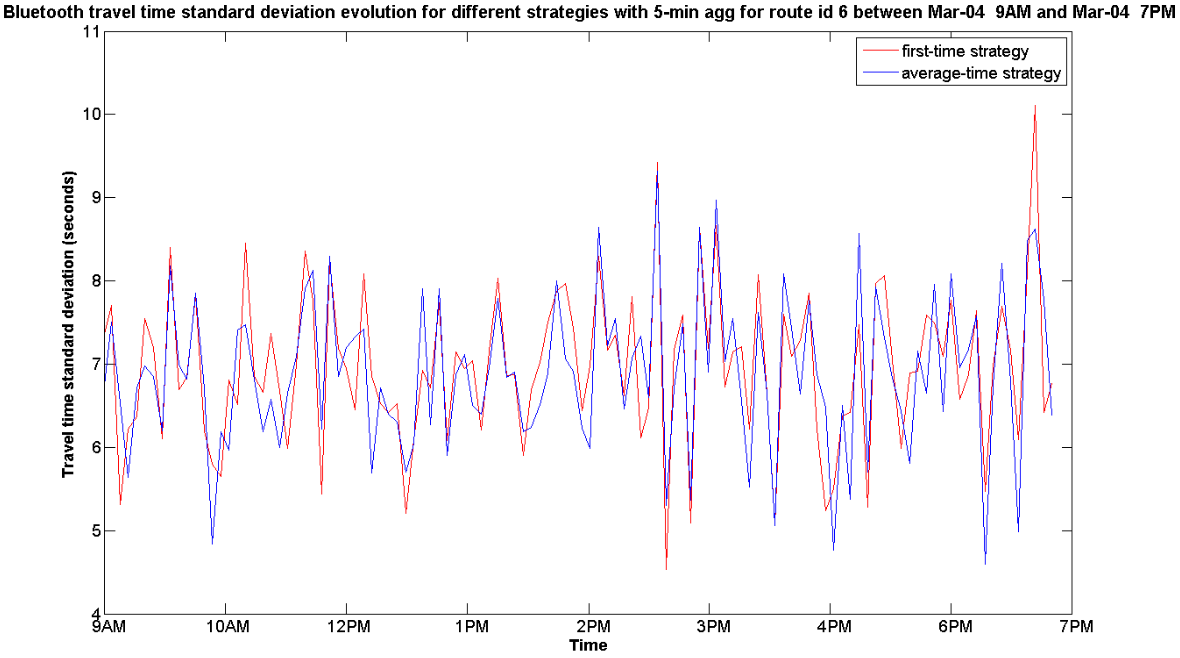

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 1 between Mar-04 9AM and Mar-04 7PM

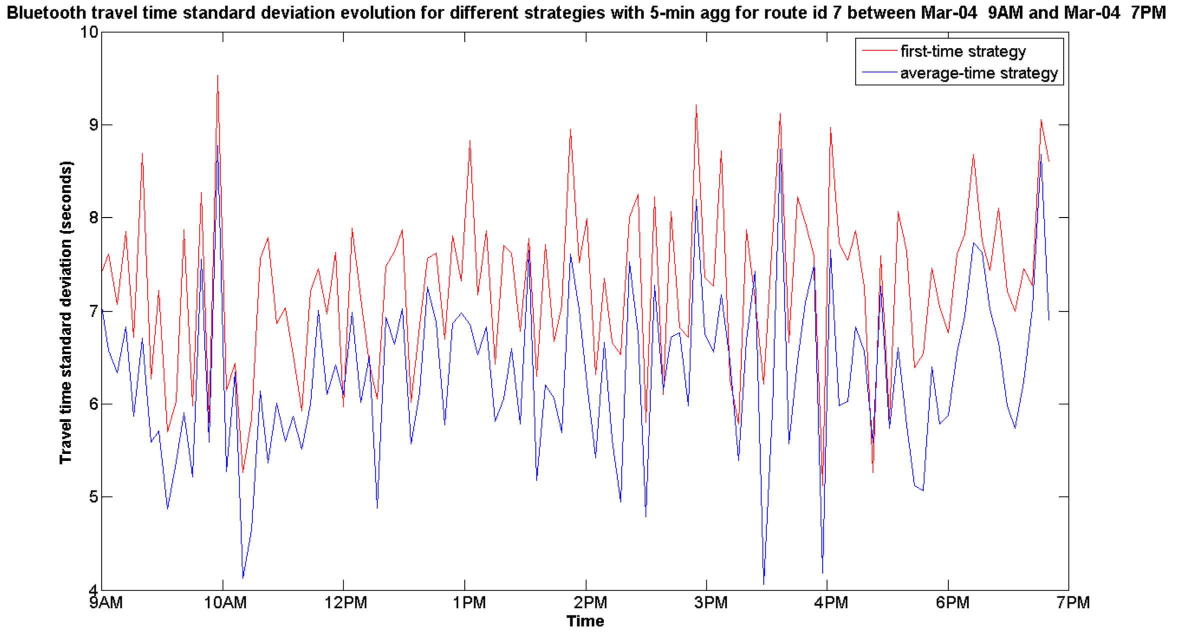

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 2 between Mar-04 9AM and Mar-04 7PM

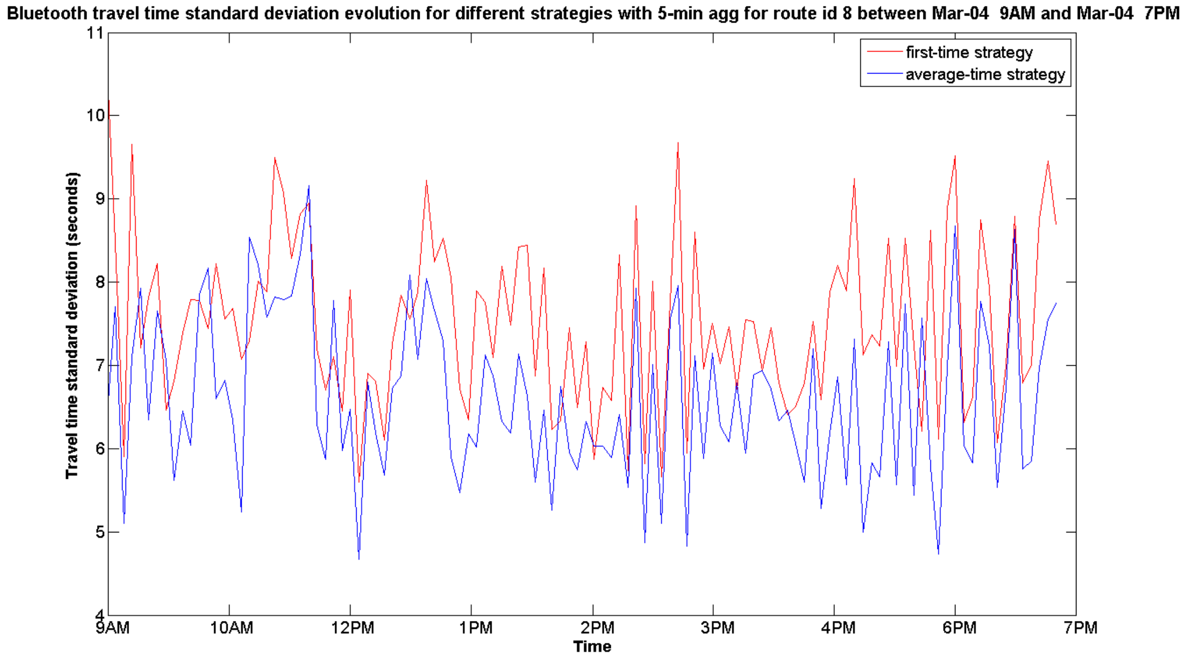

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 3 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

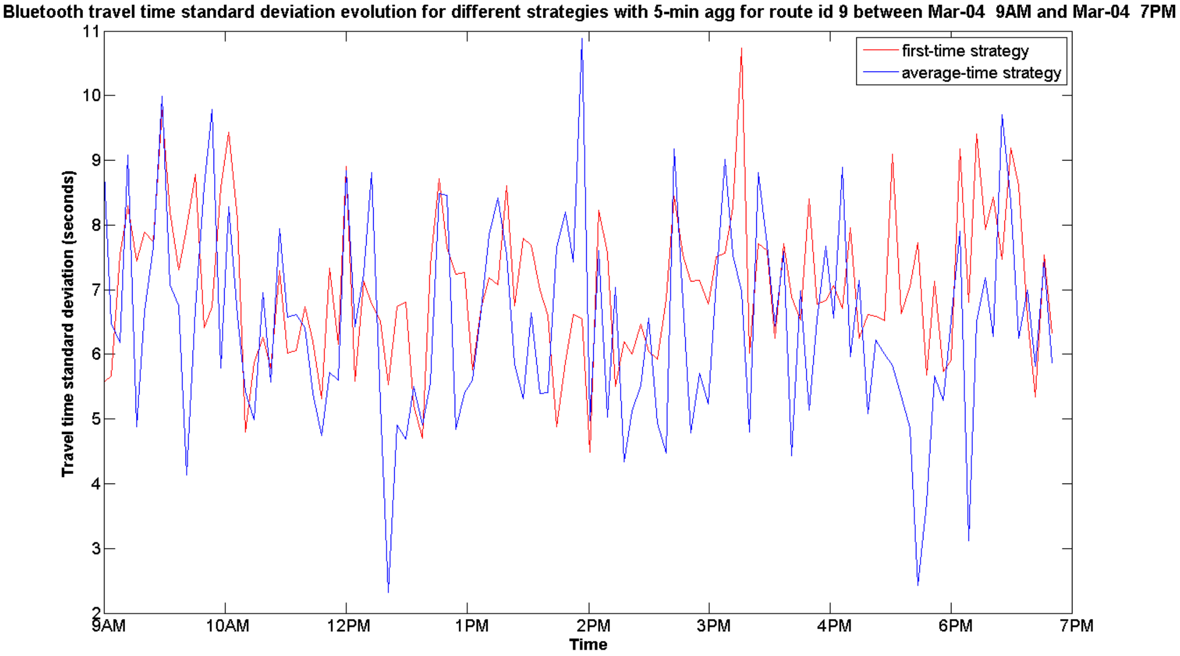

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 4 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
16
14
12
10
8
6
4
2
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time


Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 5 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
13
12
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

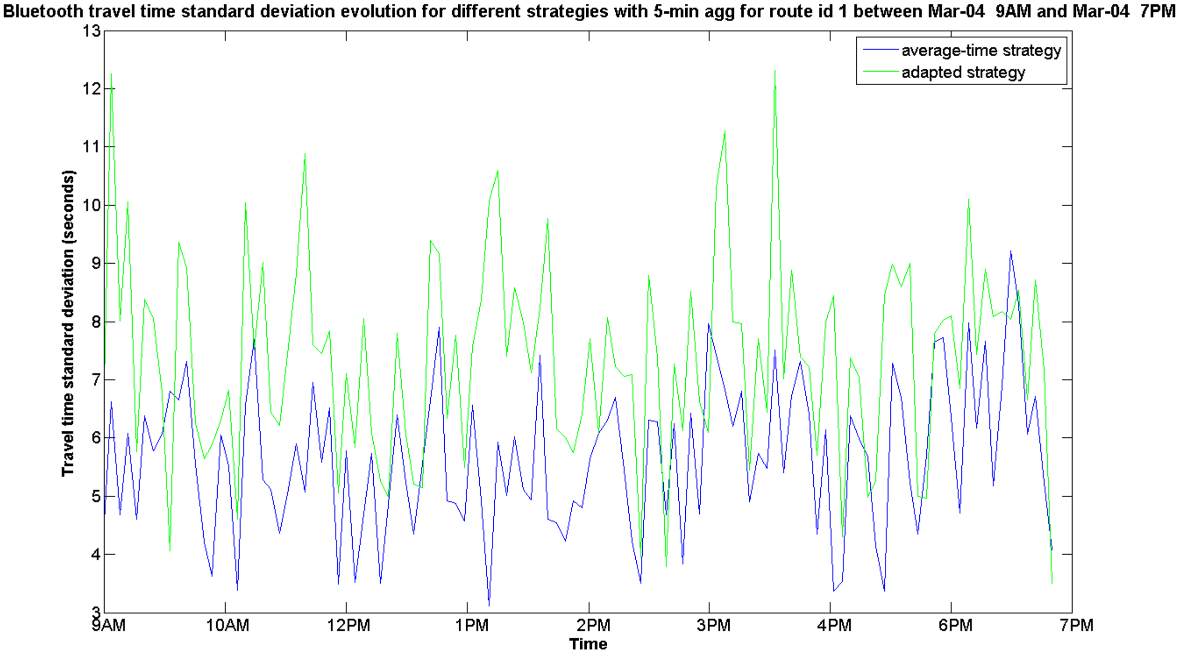

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 6 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

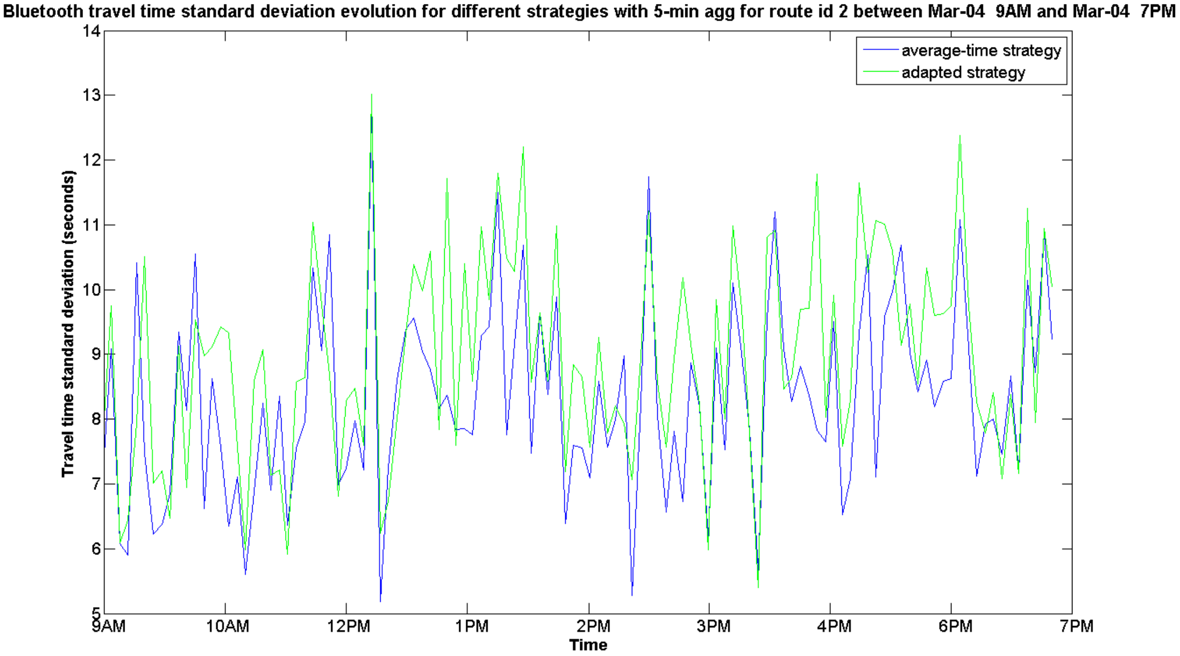

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 7 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

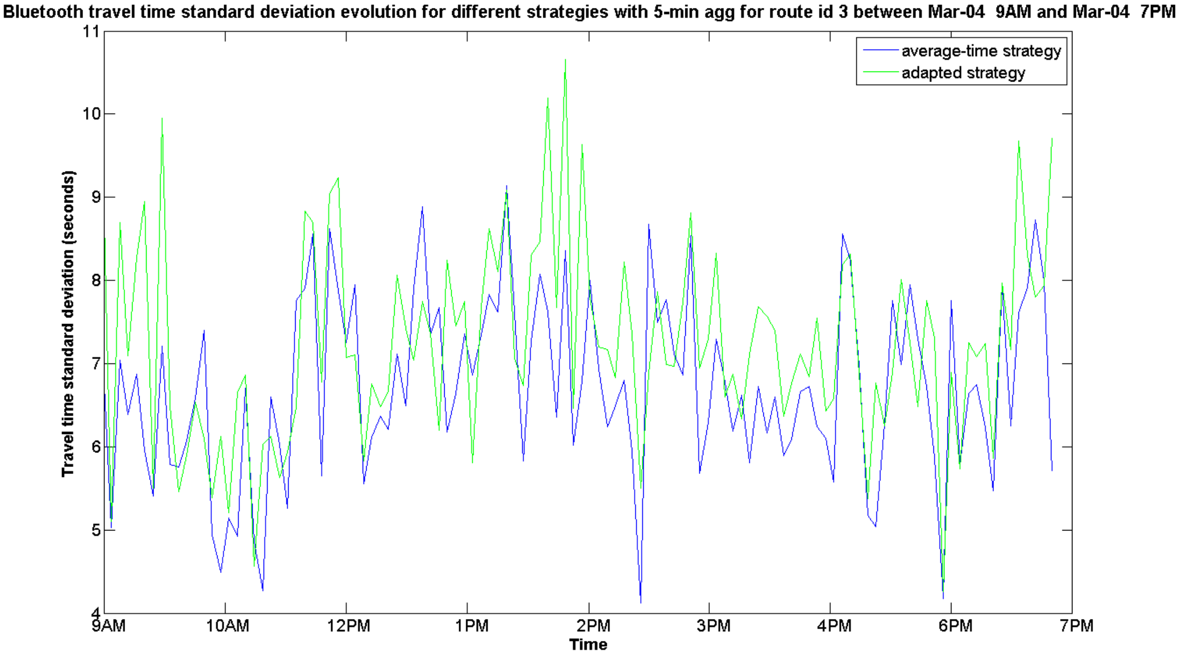

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 8 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

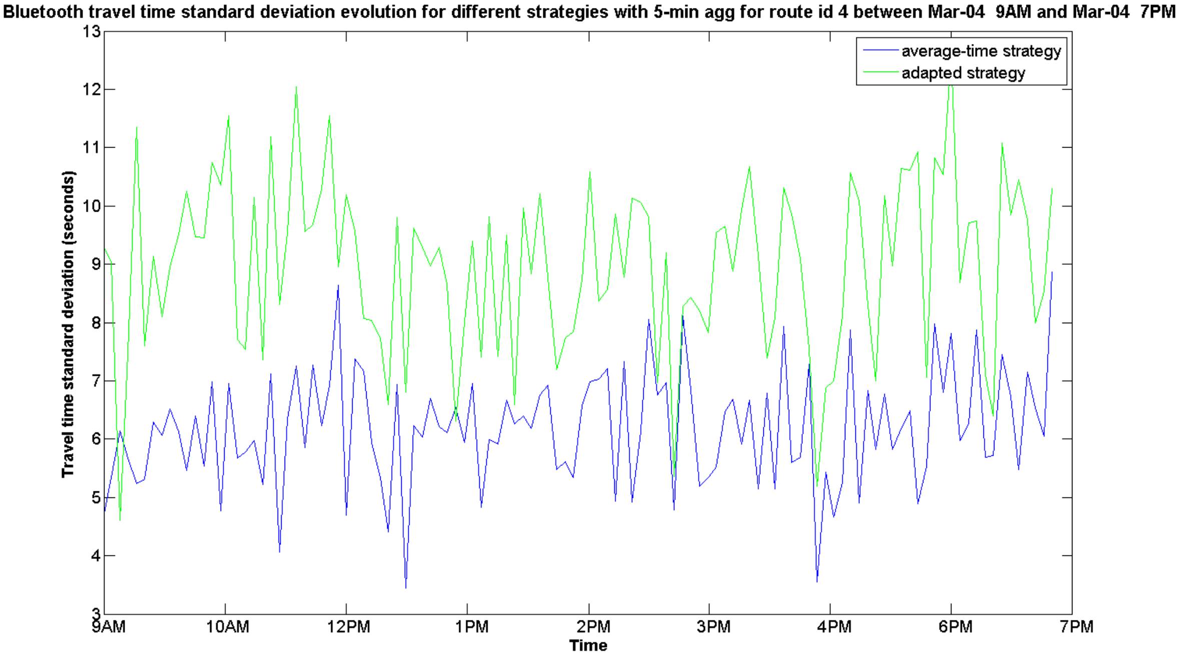

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 9 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
Travel time standard deviation (seconds)
11
10
9
8
7
6
5
4
3
2
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

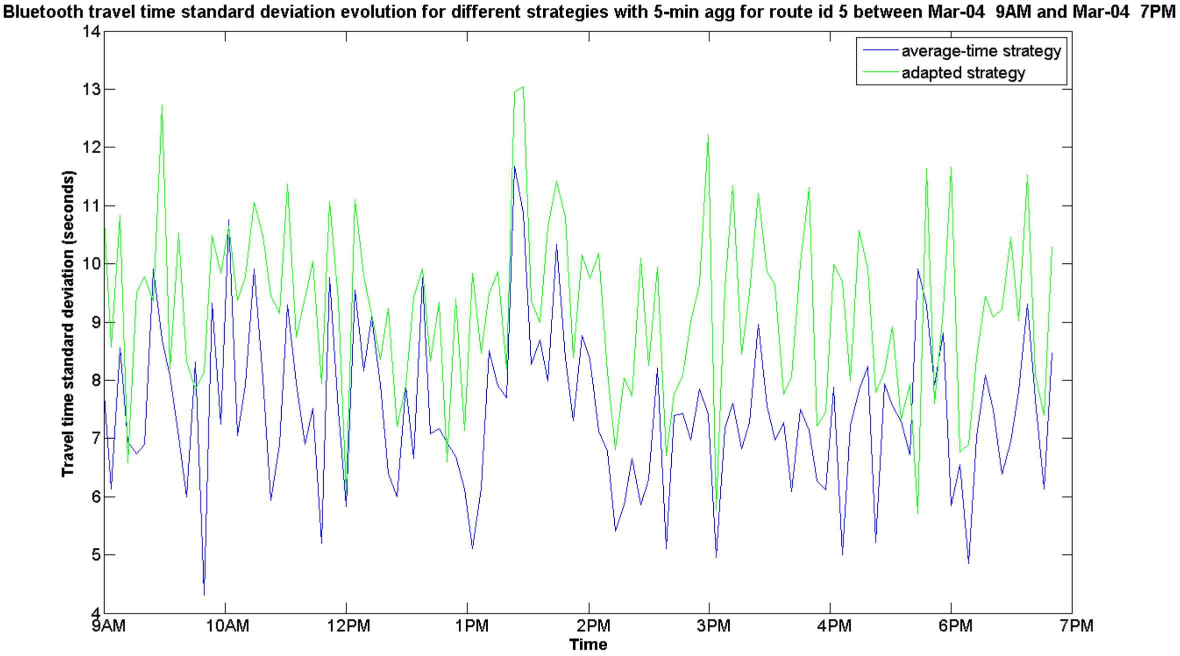
“Average time” versus “Adapted”

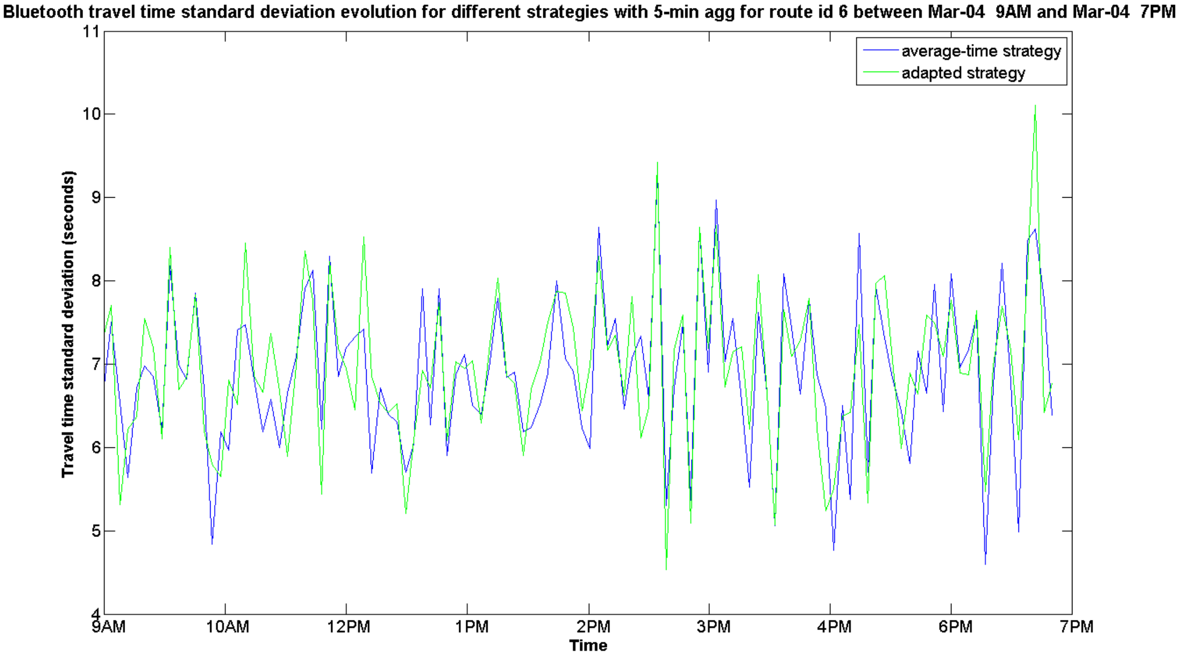

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 1 between Mar-04 9AM and Mar-04 7PM

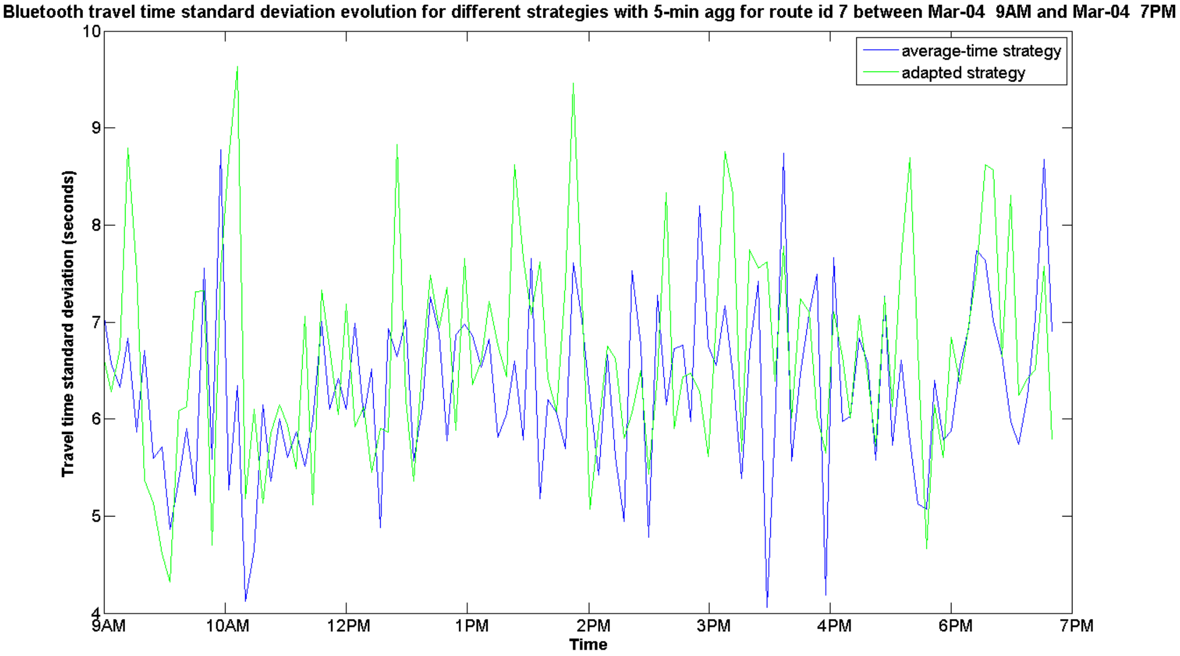

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 2 between Mar-04 9AM and Mar-04 7PM

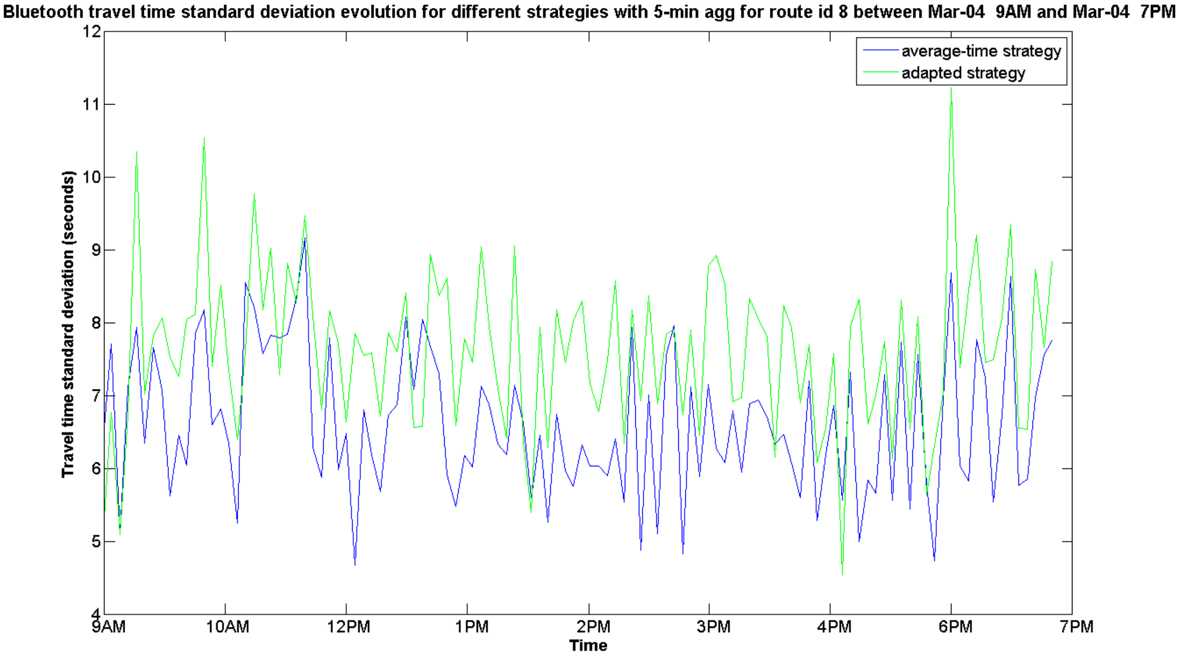

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 3 between Mar-04 9AM and Mar-04 7PM
Travel time standard deviation (seconds)
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time
average-time strategy
adapted strategy

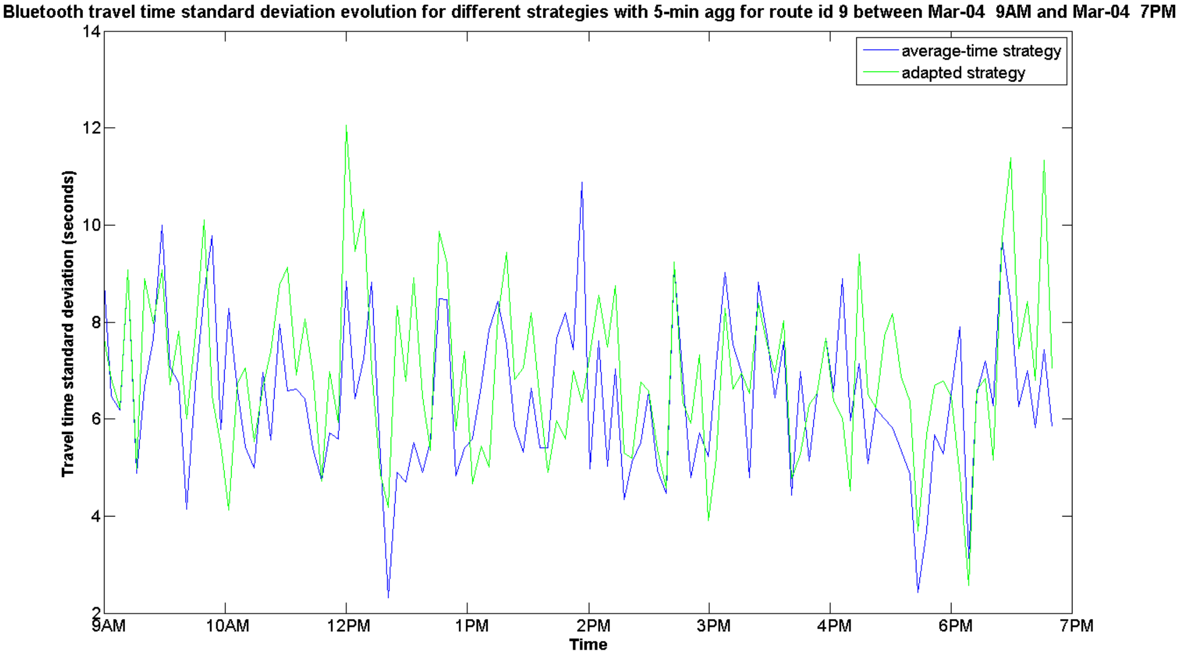

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 4 between Mar-04 9AM and Mar-04 7PM
Travel time standard deviation (seconds)
13
12
11
10
9
8
7
6
5
4
3
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time
average-time strategy
adapted strategy


Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 5 between Mar-04 9AM and Mar-04 7PM
Travel time standard deviation (seconds)
14
13
12
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time
average-time strategy
adapted strategy


Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 6 between Mar-04 9AM and Mar-04 7PM
Travel time standard deviation (seconds)
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time
average-time strategy
adapted strategy

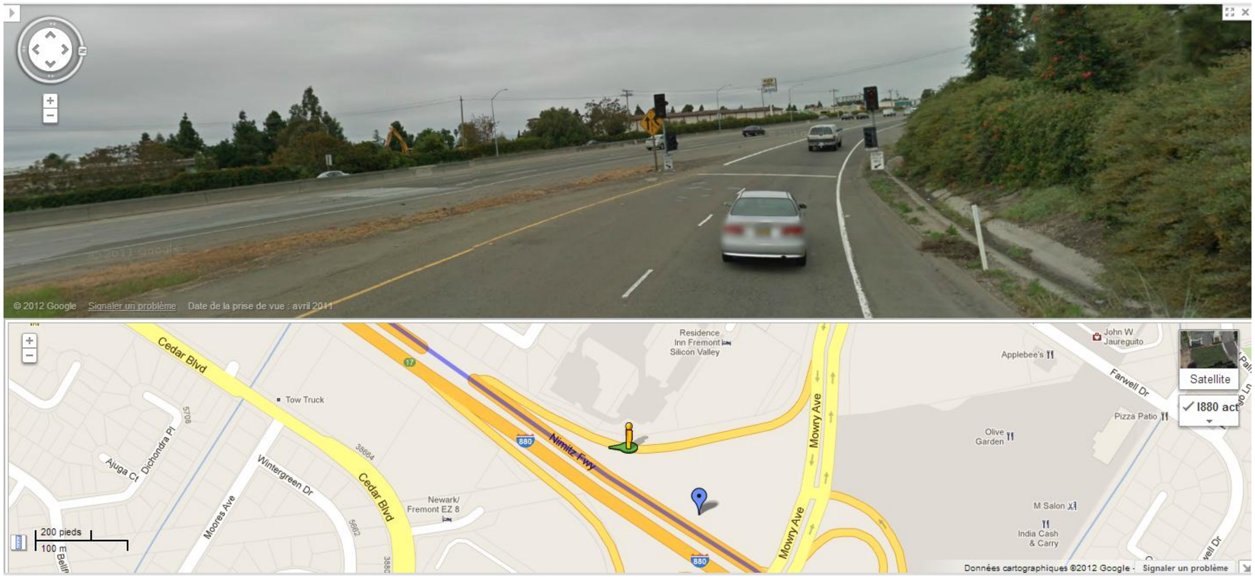

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 7 between Mar-04 9AM and Mar-04 7PM
average-time strategy
adapted strategy
Travel time standard deviation (seconds)
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time


Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 8 between Mar-04 9AM and Mar-04 7PM
average-time strategy
adapted strategy
Travel time standard deviation (seconds)
12
11
10
9
8
7
6
5
4
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

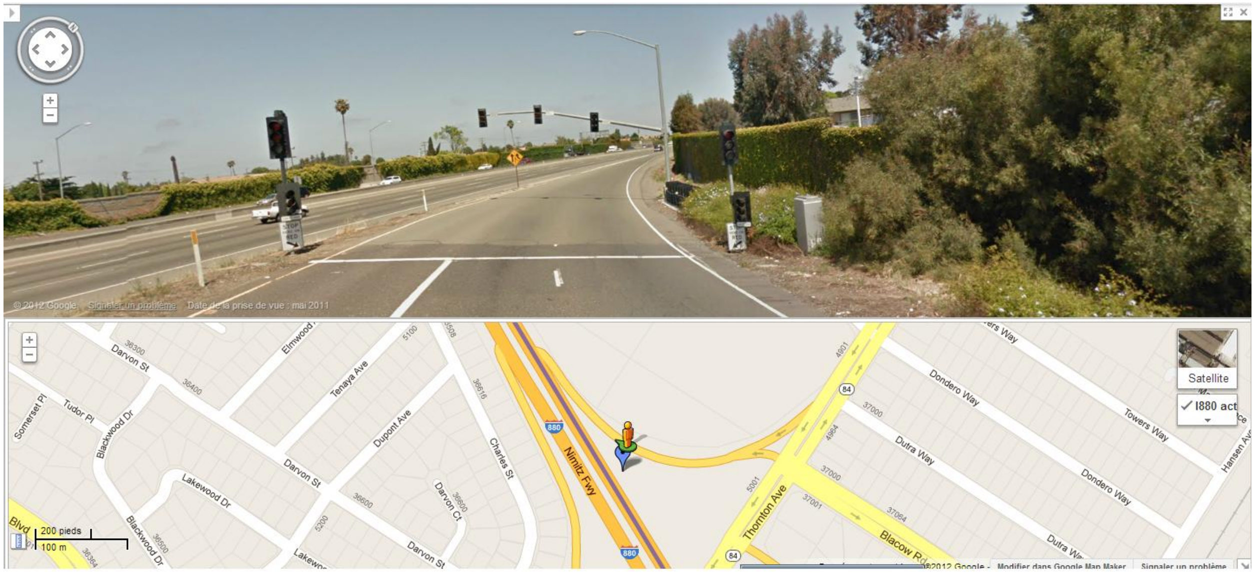

Bluetooth travel time standard deviation evolution for different strategies with 5-min agg for route id 9 between Mar-04 9AM and Mar-04 7PM
average-time strategy
adapted strategy
Travel time standard deviation (seconds)
14
12
10
8
6
4
2
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

3.5 ROAD ENVIRONMENTS OF BLUETOOTH SENSORS ON I-880 NORTHBOUND

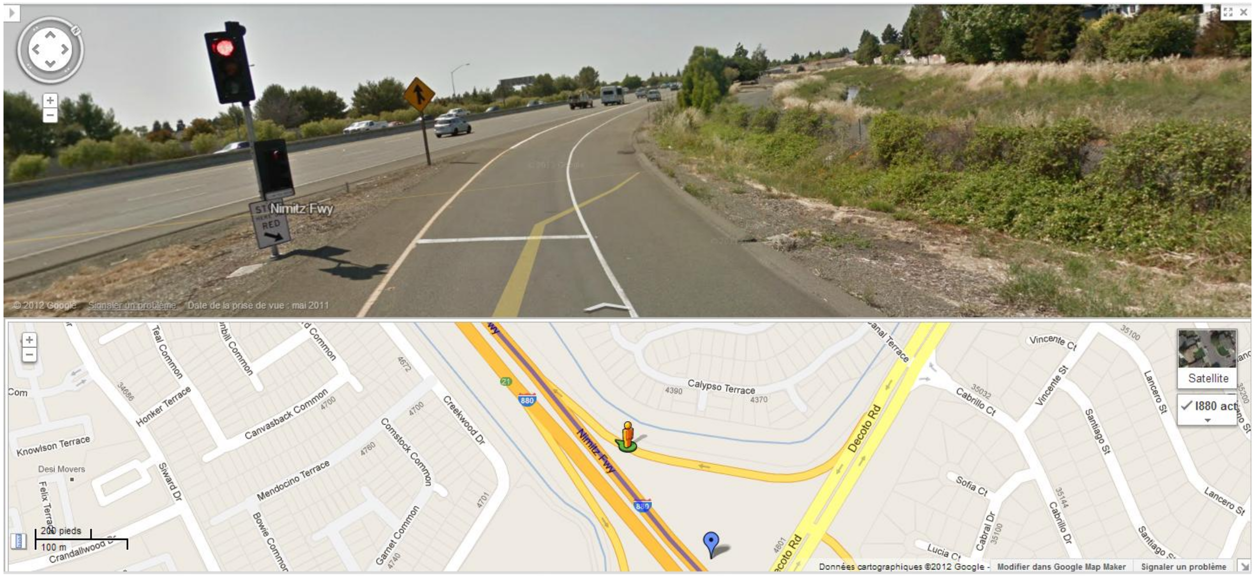
On-ramp and metered light at detector 1



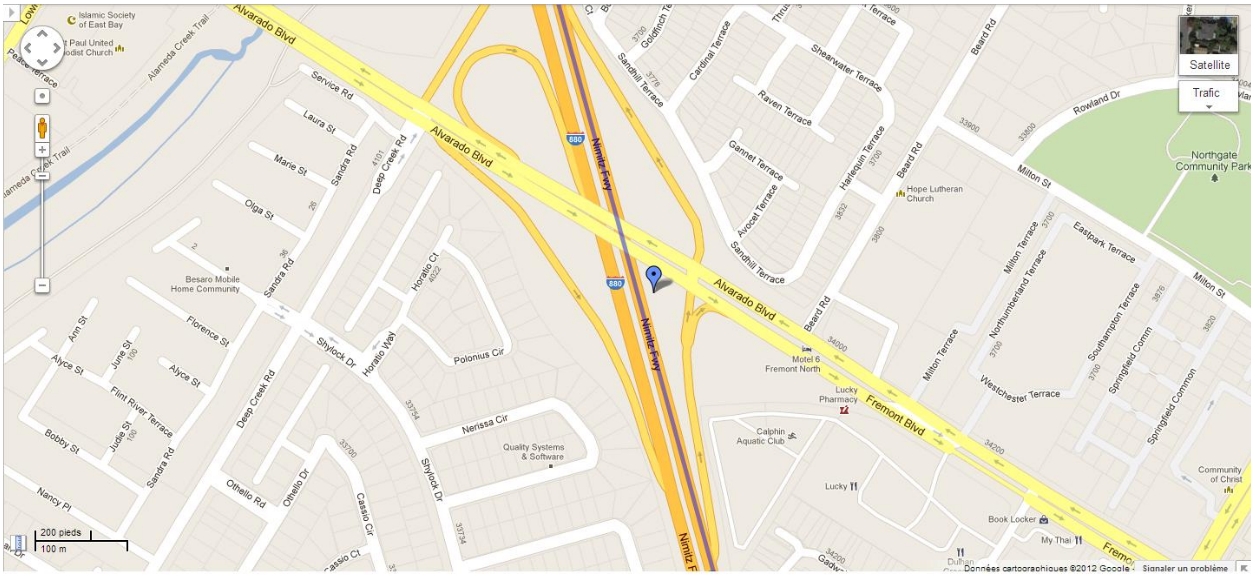
On-ramp and metered light detector 2



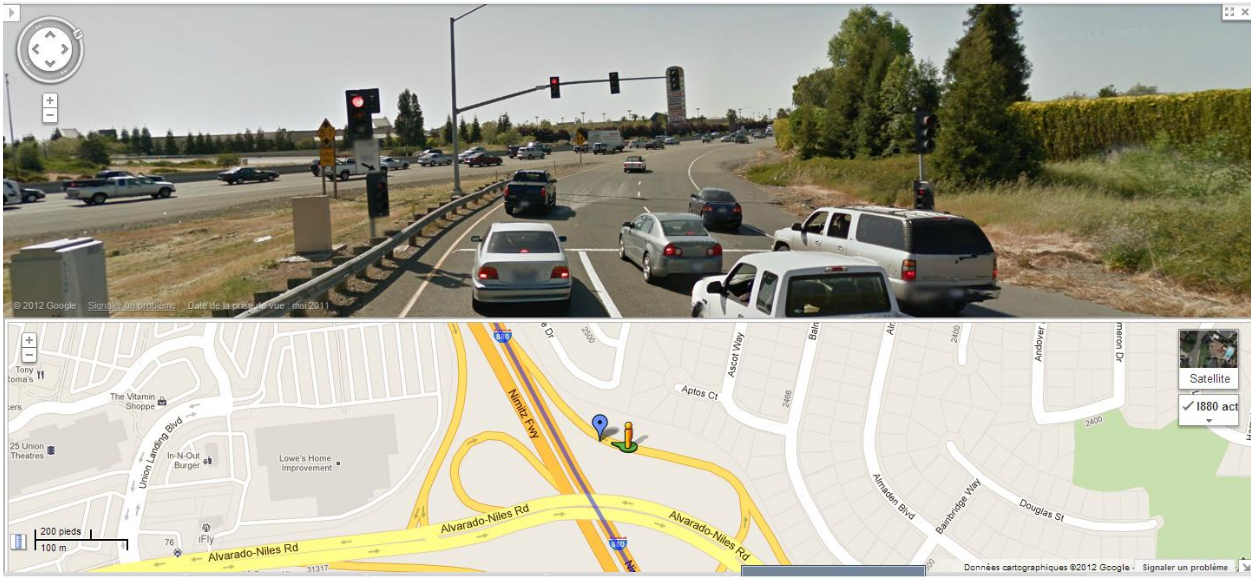
On-ramp and metered light detector 3



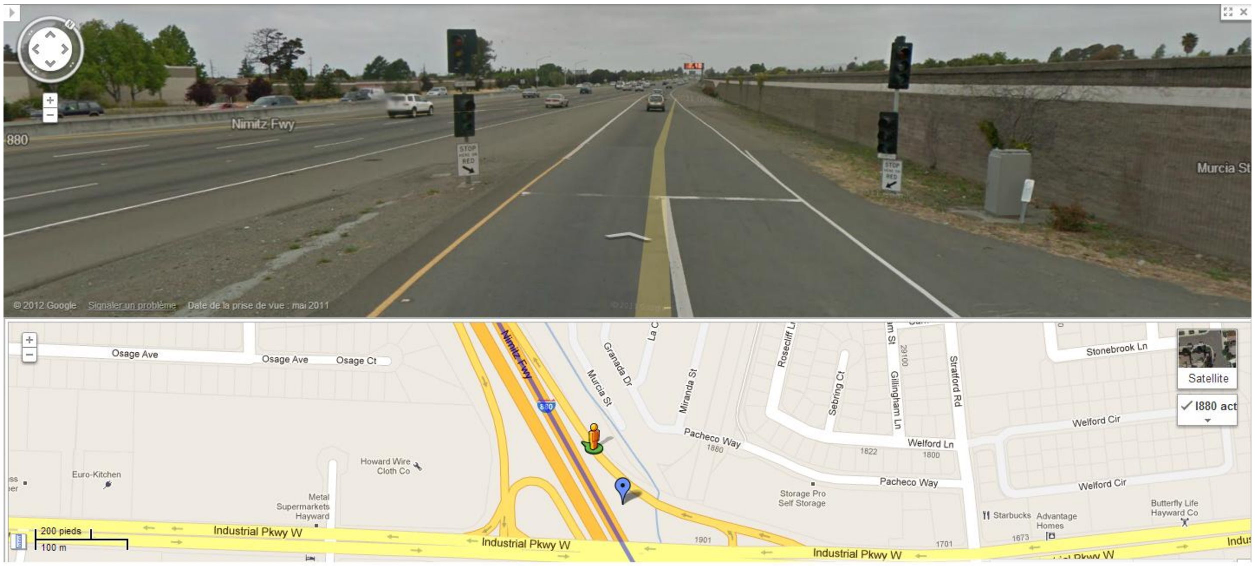
On-ramp and metered light detector 4



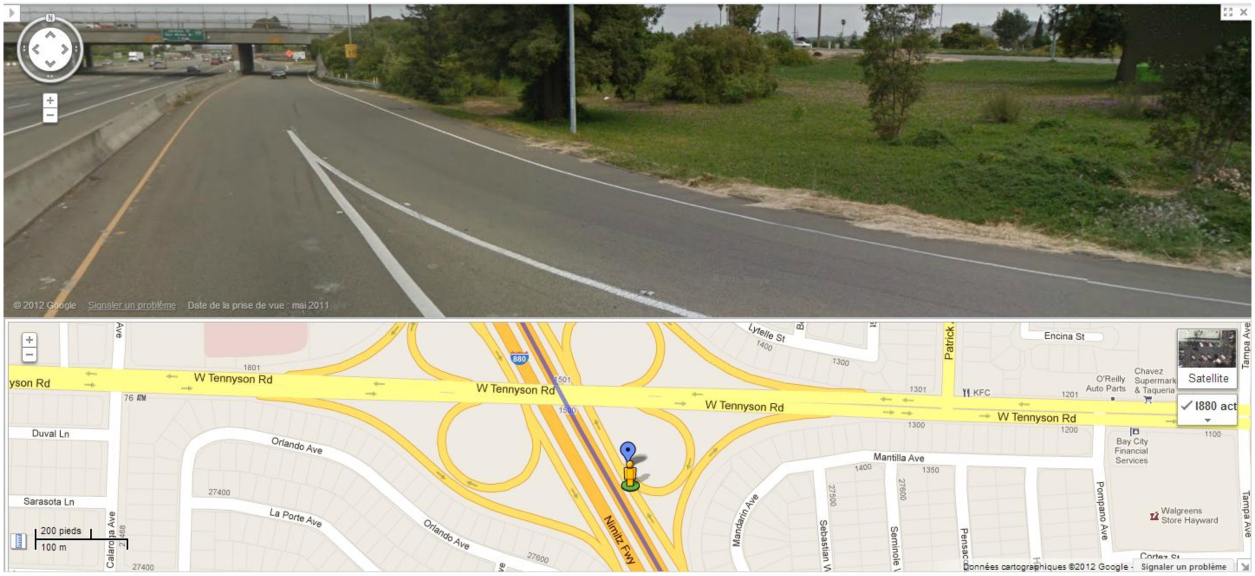
Detector 5



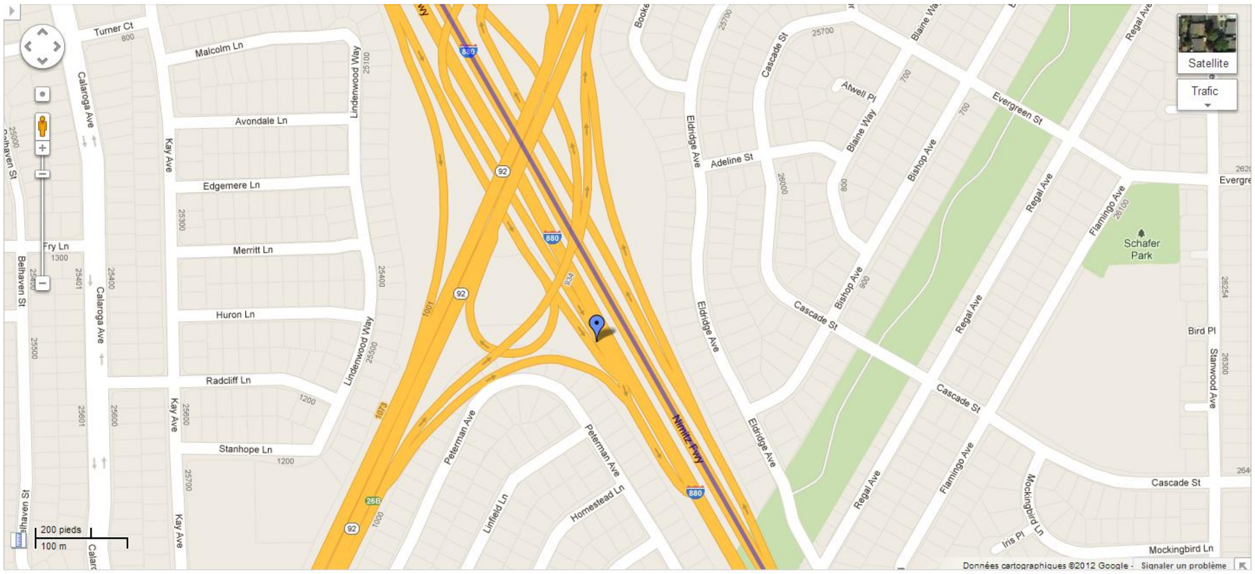
On-ramp and metered light at detector 6



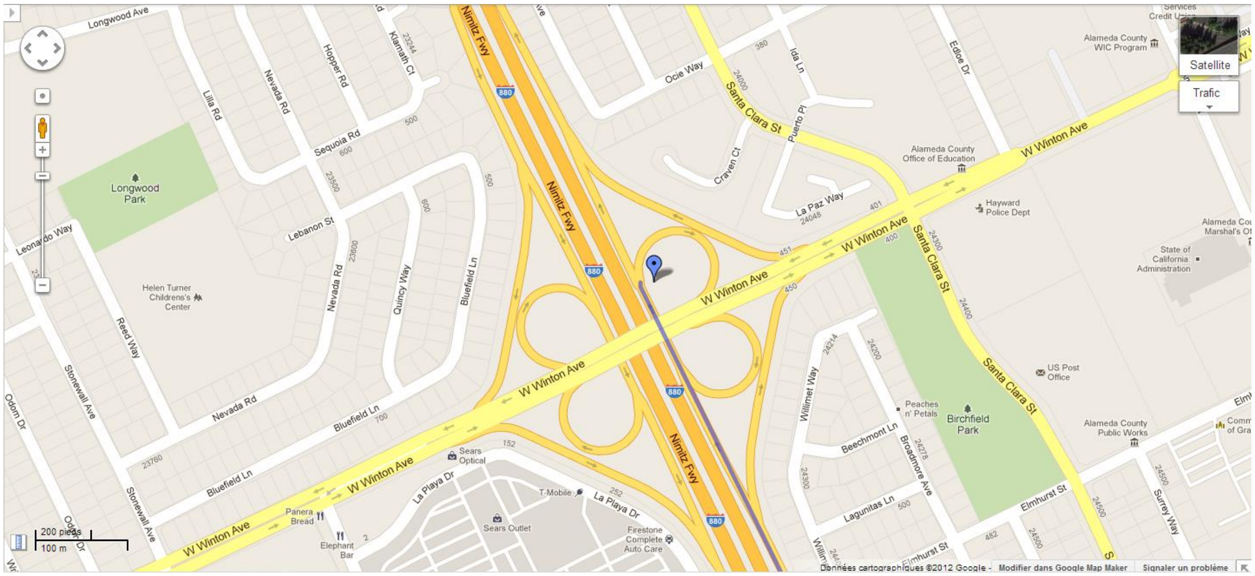
On-ramp and metered light detector 7



Detector 8

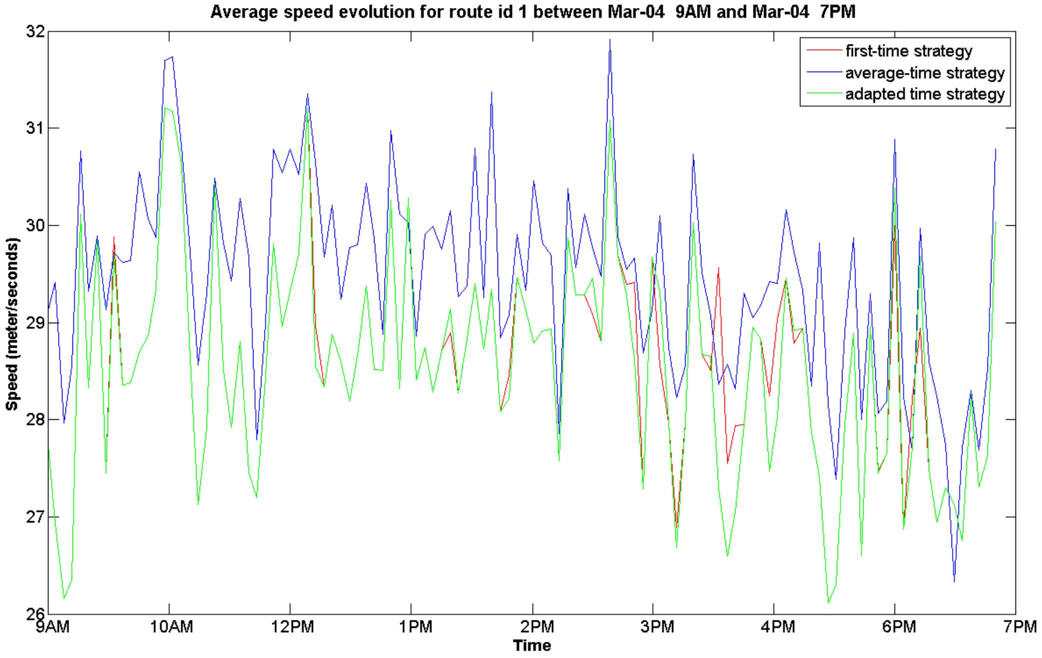


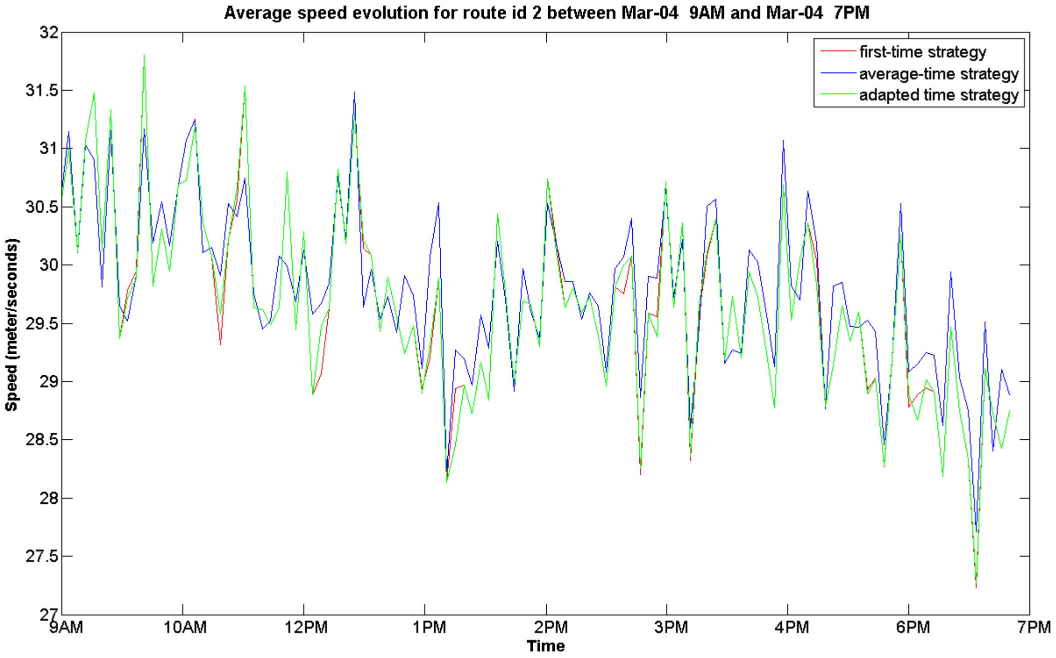
Detector 9

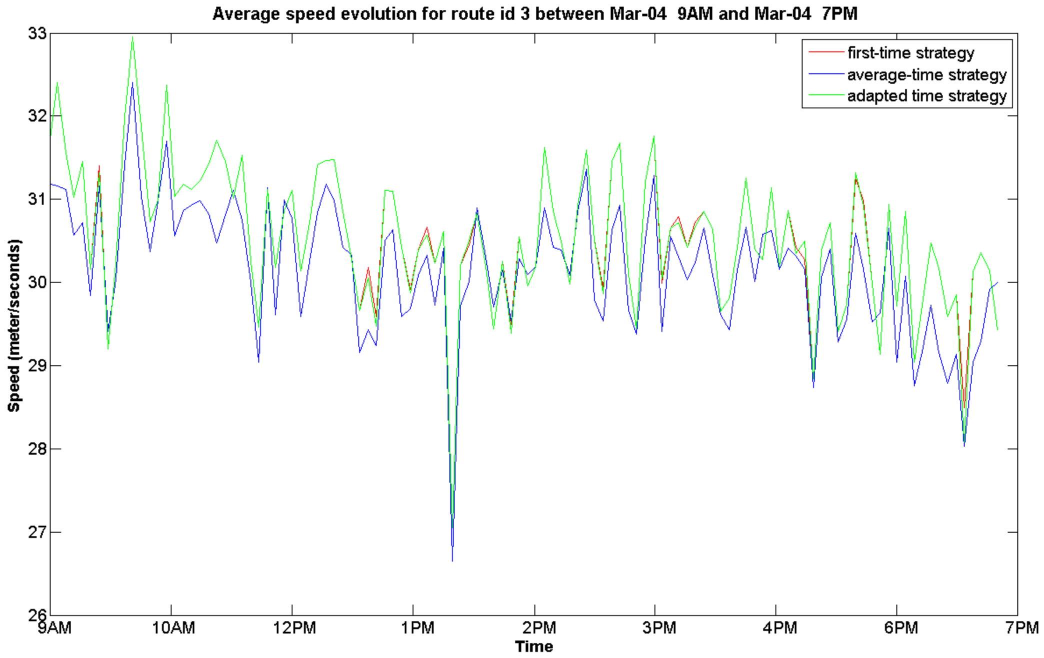


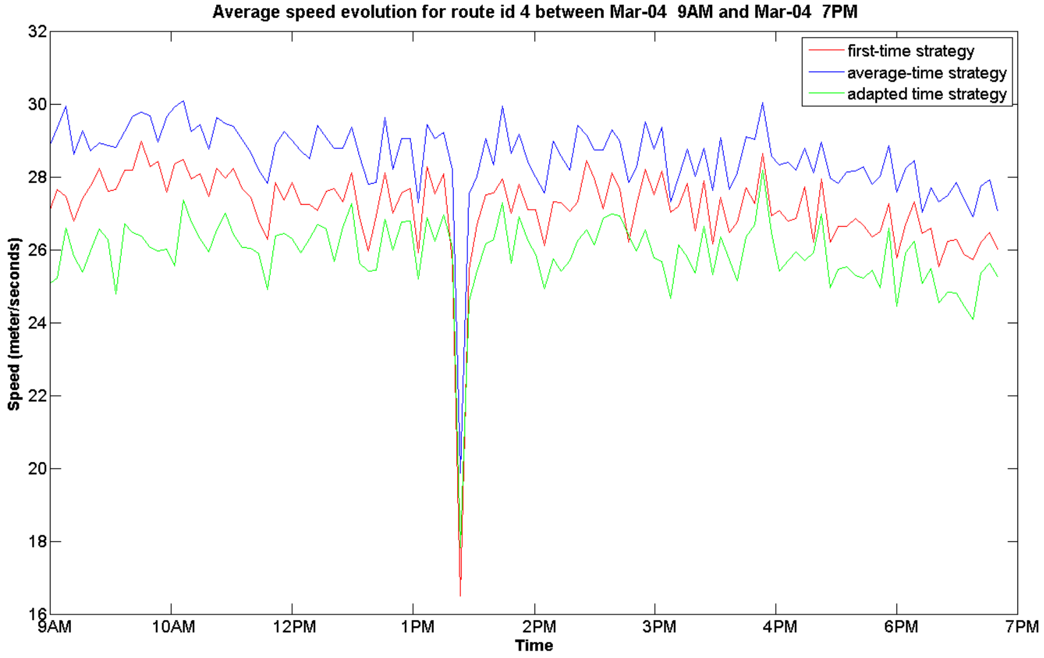
Detector 10

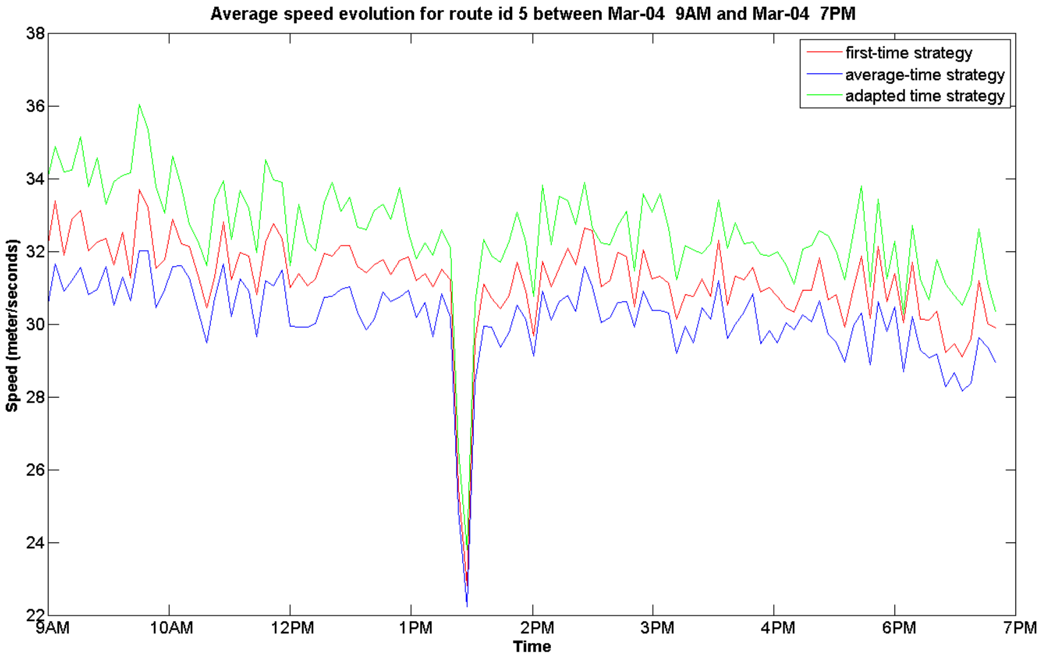
3.6 AVERAGE SPEED EVOLUTION FOR ALL ROUTES AND ALL STRATEGIES

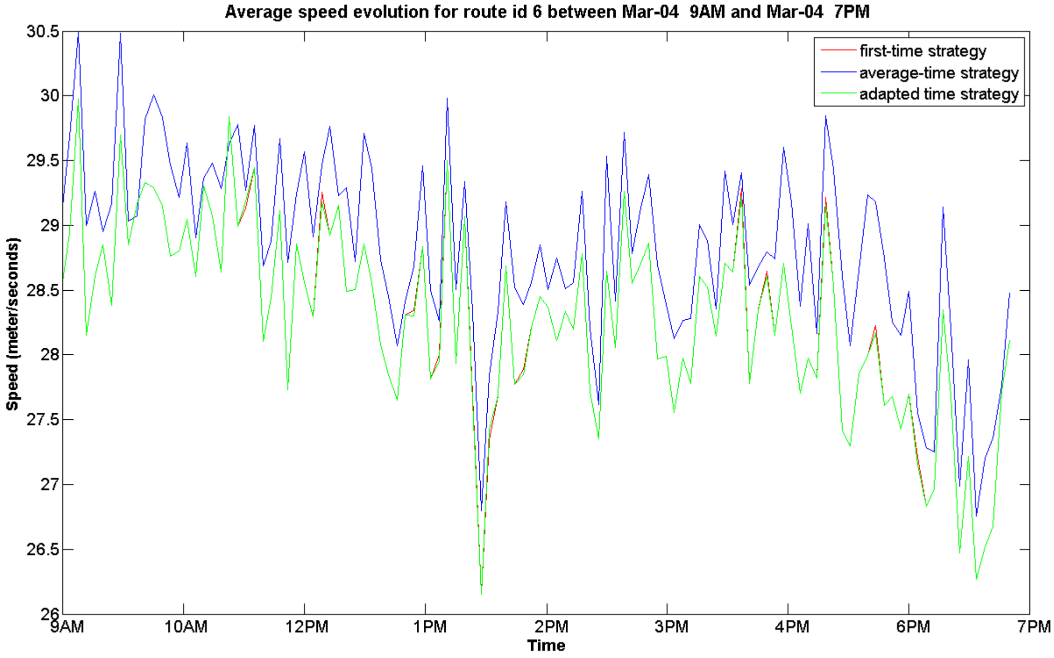


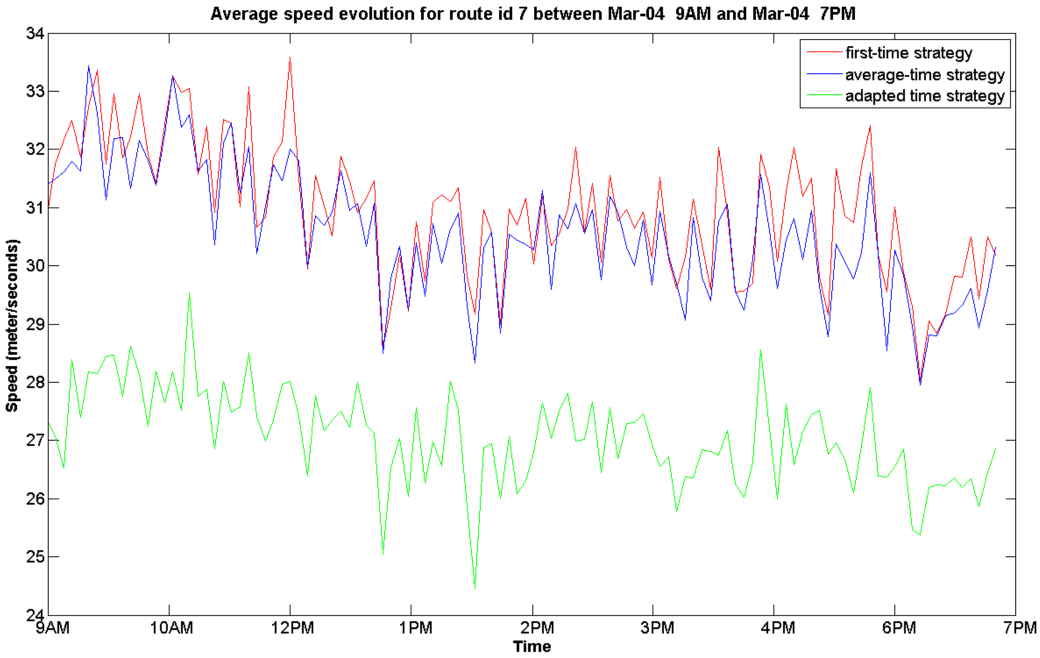


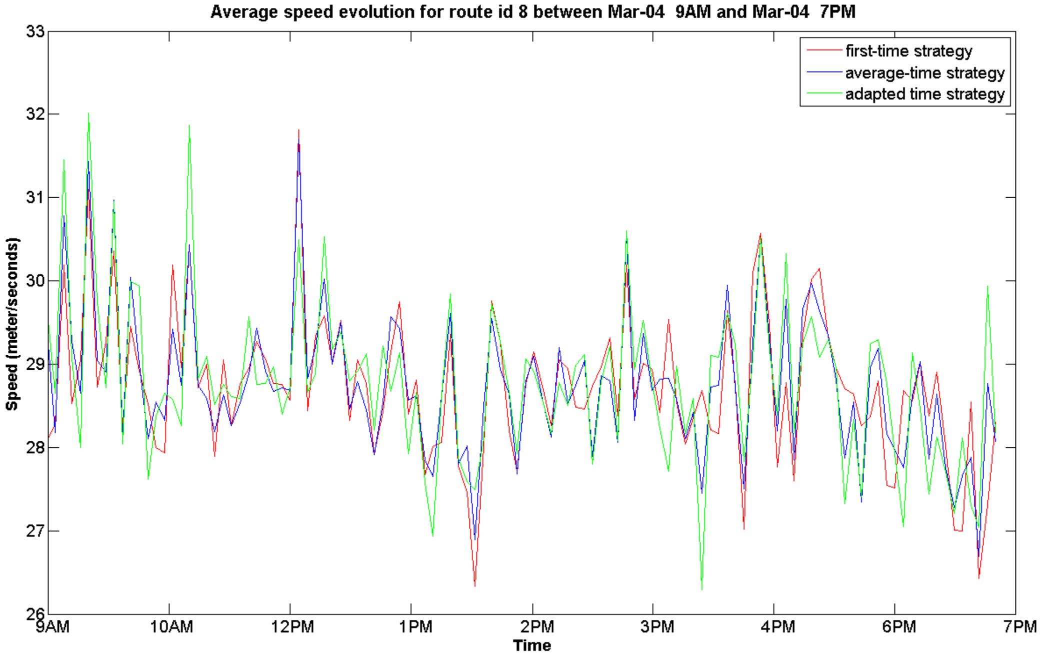

Average speed evolution for route id 3 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
adapted time strategy
Speed (meter/seconds)
33
32
31
30
29
28
27
26
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

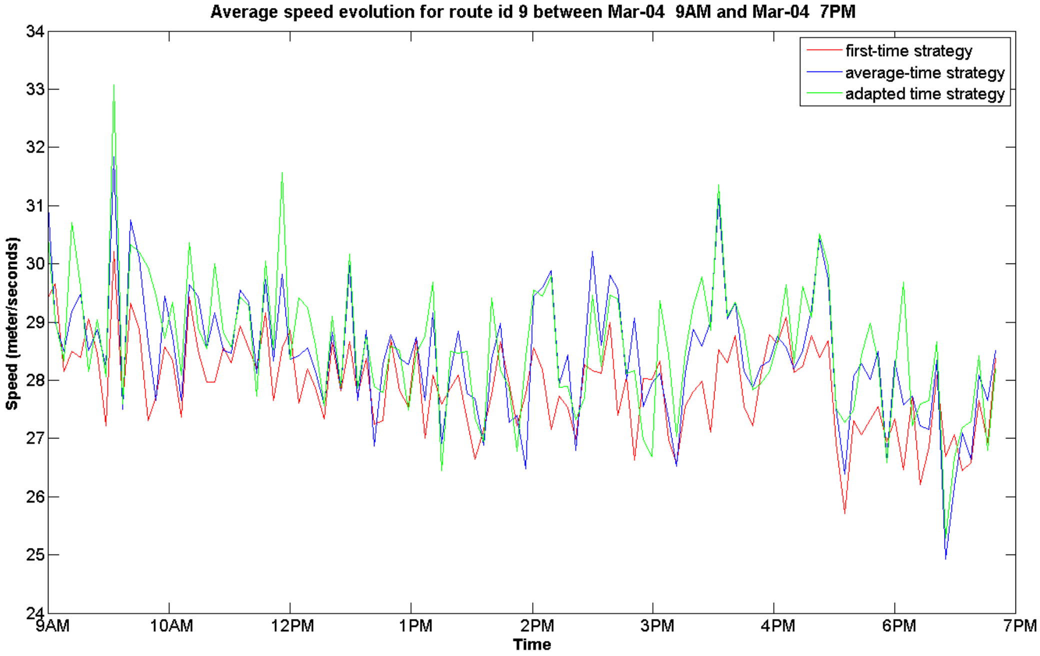

Average speed evolution for route id 4 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
adapted time strategy
Speed (meter/seconds)
32
30
28
26
24
22
20
18
16
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time


Average speed evolution for route id 5 between Mar-04 9AM and Mar-04 7PM
Speed (meter/seconds)
38
36
34
32
30
28
26
24
22
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time
first-time strategy
average-time strategy
adapted time strategy

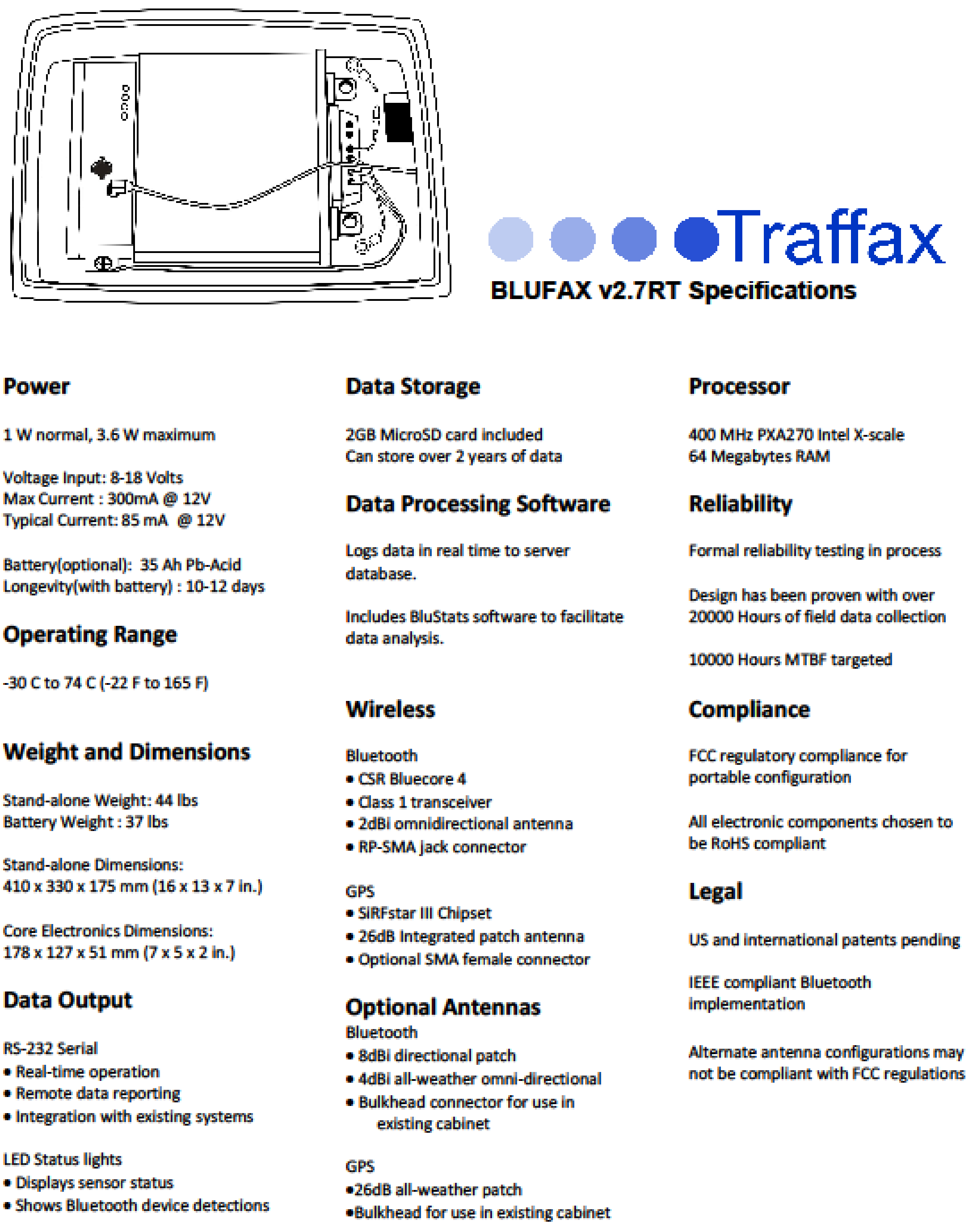

Average speed evolution for route id 6 between Mar-04 9AM and Mar-04 7PM
Speed (meter/seconds)
30.5
30
29.5
29
28.5
28
27.5
27
26.5
26
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time
first-time strategy
average-time strategy
adapted time strategy


Average speed evolution for route id 7 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
adapted time strategy
Speed (meter/seconds)
34
33
32
31
30
29
28
27
26
25
24
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time


Average speed evolution for route id 8 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
adapted time strategy
Speed (meter/seconds)
33
32
31
30
29
28
27
26
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time


Average speed evolution for route id 9 between Mar-04 9AM and Mar-04 7PM
first-time strategy
average-time strategy
adapted time strategy
Speed (meter/seconds)
34
33
32
31
30
29
28
27
26
25
24
9AM
10AM
12PM
1PM
2PM
3PM
4PM
6PM
7PM
Time

4 BLUETOOTH SENSOR SPECIFICATIONS



Traffax

BLUFAX v2.7RT Specifications

Power

1 W normal, 3.6 W maximum

Voltage Input: 8-18 Volts
Max Current : 300mA @ 12V
Typical Current: 85 mA @ 12V

Battery(optional): 35 Ah Pb-Acid
Longevity(with battery) : 10-12 days

Operating Range

-30 C to 74 C (-22 F to 165 F)

Weight and Dimensions

Stand-alone Weight: 44 lbs
Battery Weight : 37 lbs

Stand-alone Dimensions:
410 x 330 x 175 mm (16 x 13 x 7 in.)

Core Electronics Dimensions:
178 x 127 x 51 mm (7 x 5 x 2 in.)

Data Output

RS-232 Serial
- Real-time operation
- Remote data reporting
- Integration with existing systems

LED Status lights
- Displays sensor status
- Shows Bluetooth device detections

Data Storage

2GB MicroSD card included
Can store over 2 years of data

Data Processing Software

Logs data in real time to server database.

Includes BluStats software to facilitate data analysis.

Wireless

Bluetooth
- CSR Bluecore 4
- Class 1 transceiver
- 2dBi omnidirectional antenna
- RP-SMA jack connector

GPS
- SiRFstar III Chipset
- 26dB Integrated patch antenna
- Optional SMA female connector

Optional Antennas

Bluetooth
- 8dBi directional patch
- 4dBi all-weather omni-directional
- Bulkhead connector for use in existing cabinet

GPS
- 26dB all-weather patch
- Bulkhead for use in existing cabinet

Processor

400 MHz PXA270 Intel X-scale
64 Megabytes RAM

Reliability

Formal reliability testing in process

Design has been proven with over 20000 Hours of field data collection

10000 Hours MTBF targeted

Compliance

FCC regulatory compliance for portable configuration

All electronic components chosen to be RoHS compliant

Legal

US and international patents pending

IEEE compliant Bluetooth implementation

Alternate antenna configurations may not be compliant with FCC regulations

Chapter 4

Procuring Traffic Data

After interviewing stakeholders, analyzing previous efforts to acquire probe data, choosing the sites we wanted data for, and establishing methods for estimating ground truth for those sites, we were ready to test the waters of the probe data market by creating and issuing our own Request for Proposal (RFP). This chapter describes the process we went through to develop an RFP, decide what to ask for, identify the components to include, and evaluate the responses. The chapter concludes with recommendations to Caltrans based on our experience.

1 PROCURING DATA: THE RFP PROCESS

A pivotal step in procuring traffic data from private vendors is the Request for Proposal (RFP) process: defining the data you want, soliciting proposals from potential suppliers, reviewing responses, selecting a vendor, and contracting for product delivery. As part of its effort to assess various methods for procuring and fusing traffic data from multiple sources, PATH[3] issued a request for proposals on May 9, 2011 for the provision of point-speed data over a three-month period. The purpose of the RFP was to explore the feasibility of purchasing unaggregated data, the existing market for such data, its market price, coverage, and quality. (For the complete text of the RFP, see the Supporting Documents for Task Orders 1 and 2, provided to Caltrans in electronic format along with this report.)

This chapter summarizes the main components of PATH's RFP and makes recommendations based on what we learned. It identifies data-type candidates for procurement, data sources and vendors, and methodologies for obtaining data, and addresses these key parts of the process:

- **Data specification**—What information do we want?
- **Data collection**—How is it gathered, where does it come from, what does it cover?
- **Data presentation and definition of coverage metrics**—How should vendors present the data? How granular should it be?
- **Data quality**—What constitutes data quality?
- **Purchasing and pricing**—What are the purchasing and pricing options?
- **RFP characteristics**—What makes a good RFP?
- **Proposal evaluation criteria**—How do we judge a proposal and select a vendor?

---

[3] The RFP described in this chapter was issued by the California Center for Innovative Transportation (CCIT). CCIT has since merged with Partners for Advanced Transportation Technology (PATH) and operates under that name.

2 DATA SPECIFICATION: WHAT INFORMATION WE WANT

We wanted to procure unprocessed probe vehicle data, specifically point-speed data. Delivering this unprocessed data as coherent traffic information to public and transportation agencies requires an algorithm that manages relationships with probe data sources and fuses their information together effectively.

Some data suppliers provide probe data that has already been aggregated and processed into travel time information. We chose instead to procure *unprocessed* traffic data, for two reasons:

- First, Caltrans is interested in purchasing unprocessed traffic data in order to know the source of the data for validation and to have better insight into the data's quality, density, and coverage. They can then fuse it by an algorithm that is known to them to create travel time information.
- Second, PATH is developing the fusion algorithm and will process the purchased data with the Mobile Millennium highway model to create travel time information.

In June 2010, PATH issued a Request for Information (RFI) to solicit information from the vendor community. We used the results to develop a targeted data specification (spatial, temporal, quality, volume, methods of delivery). The RFI and vendor responses are included in the Supporting Documents for Task Orders 1 and 2, provided to Caltrans in electronic format along with this report.

It was important to specify unaggregated data clearly. We therefore defined one *point-speed datum* as a measurement of the location and speed of a vehicle on the highway segment of interest, consisting of at least the mandatory fields listed in Table 6. All fields marked optional were desired attributes that help PATH make the best use of the data. Even if only some data were available for the optional fields, they could still provide useful traffic information.

Table 6: Data Specifications

Data Name	Description	Format	Unit	Mandatory/ Optional
Identifier	A unique identifier for the device for which a reading is provided	Alphanumeric		Optional
Measurement Timestamp	Time of measurement	Numeric		Mandatory
Point	Latitude and longitude as reported by the sensor	Numeric	° (decimal degrees)	Mandatory
Speed	Speed of vehicle as reported by the sensor	Numeric	m/s	Mandatory
Velocity	3-D velocity vector corresponding to “Speed”	Numeric	m/s	Optional
Heading	True heading of vehicle	Numeric	° (decimal degrees)	Optional
Altitude	Altitude	Numeric	m	Optional
Vehicle Type	Type of vehicle (Vendor to describe an identification scheme; PATH has no specific requirement)	Alphanumeric		Optional
Lane Identifier	Lane number (sequentially from 1, which corresponds to the leftmost lane in the direction of interest)	Numeric		Optional

3 DATA COLLECTION: HOW IT’S GATHERED AND WHAT IT COVERS

It was important to have a well-defined description of the data collection. We asked vendors to provide the following information:

- Describe the data collection and processing methods that will be used to generate the supplied data or origin of data if data is acquired by vendor:
    - Describe relevant filtering or processing procedures used before the data is sent out.
    - Are measured locations mapped to the road network by way of a mathematical transformation?
    - How are data points selected as belonging to the given roads of interest (one segment of I-880 in Northern California and three segments of I-15 in Southern California)?
    - Are outliers removed (and how do you define outliers)? How?
- Describe the spatial coverage for the roads of interest (I-880 and I-15). Is data equally available on the entire road, or are some parts covered more thoroughly than others? Explain in detail.
- Describe the sources of data and all agreements governing that data (including any terms and conditions that would affect the ability to supply the data).
- Consider the privacy of the users of the devices from which data is collected:
    - Discuss the ability for an individual vehicle’s path to be identified from the data and any measures you would take to protect this information.

4 DATA PRESENTATION AND DEFINITION OF COVERAGE METRICS

Vendor selection depended in part on the quantity of data offered in each proposal and its distribution along the relevant highways. We had extensive discussions about how much data should be required, at what level of granularity, and how it should be presented.

There were two schools of thought about the presentation of data and the definition of coverage metrics:

- **Scientific**—This approach emphasized scientific techniques and required a more detailed total data count with specific definitions of coverage metrics for each spatiotemporal zone.
- **Business**—This approach emphasized the business aspect, with the intent of stimulating the market to respond. For pragmatic reasons, this approach was open to some interpretation and did not ask for too much sensitive information.

Ultimately, we felt that the RFP required a balance of scientific rigor and simplicity—specific enough to get the data we wanted but not so complex that it discouraged vendors from responding. Consequently, we opted for the business approach and presentation of data.

4.1 BUSINESS APPROACH AND PRESENTATION OF DATA

In this approach, the RFP required vendors to provide baseline information about the quantity and distribution of their data for pre-defined segments of highways I-880 and I-15. For each segment, metrics must be calculated independently, and bidders were asked to provide average total number of point-speed counts over specific time intervals for weekdays and weekends separately. This information was to be provided on the basis of the duration of the contract (nominally three months) in Table 7.

Table 7: Sample Data Coverage Table

Highway Segment	Day of Week	Time Interval				Estimated Price
		Morning	Midday	Evening	Night	
I-880 #1	Weekday					
	Weekend					
I-880 #2	Weekday					
	Weekend					
I-15 #1	Weekday					
	Weekend					
I-15 #2	Weekday					
	Weekend					
I-15 #3	Weekday					
	Weekend					
I-15 #4	Weekday					
	Weekend					
I-15 #5	Weekday					
	Weekend					
I-15 #6	Weekday					
	Weekend					

.

The empty cells in columns labeled “Morning” through “Night” must be filled in with the average total number of points measured for the corresponding segment, day, and time interval. For example, the top-left empty cell should contain the average total number of points measured on segment #1 of I-880 between 05:00 and 09:00 for a weekday day.

5 DATA QUALITY, PURCHASING, AND PRICING

Data quality, purchasing, and pricing are, of course, essential elements of the procurement process. We had a number of discussions about these core issues, as well as what constitutes a good RFP.

5.1 DATA QUALITY

The quality of any data feed would need to be judged along the following dimensions:

- **Accuracy**—Degree to which the data from data feed matches ground truth
- **Accessibility**—Relative ease in the use of the data feed for a given traffic application
- **Completeness**—Degree to which values are present for all the fields of the data feed
- **Coverage**—Degree to which the data feed represents the whole of what is to be measured
- **Purity**—Degree to which the data feed consists of raw data
- **Validity**—Degree to which the data feed fields satisfy acceptance requirements

Beyond these general properties, however, commonly accepted methods for measuring data quality do not yet exist, and our work involved trying to understand and recommend these methods. Our discussions focused on three key considerations:

- Determining which vendor to buy from after the RFPs came in. This was complicated and relied on a number of factors, including company information, company experience, data information (quality, coverage, amount), and overall price. The implication was that assessing data quality for purposes of procurement is not just about measuring data information; it is about a lot more.

- Determining how to evaluate data we would buy during the three months we would get it. This required more in-depth analysis, fusion of data from multiple sources, and mapping to our chosen application (highway model).

- Longer-term data quality—new applications, measurements in real time, real-time fusion, etc.

5.2 CHARACTERISTICS OF A GOOD RFP

The purpose of our RFP was, in many respects, exploratory. We were testing the waters by surveying the market, seeking out a broad response, and making recommendations. Consequently, we didn't want prospective responders looking at the RFP and saying it was too complicated.

We concluded that there are several key attributes of a good RFP for procuring data, and their relative importance varies with the purpose, publicity, and value of the purchase:

- It asks for what we want: unaggregated probe data.
- It provides a way of determining the quality (appropriateness) of the product to be purchased in a way that can be compared to others.
- It stimulates the market to respond: It is easy to respond to, open to some interpretation, and does not ask for too much sensitive information. We need to think about how this data could be used to destroy the credibility of an organization if used badly. For pragmatic reasons we need to be considerate of the risks our responders (hopefully, partners) are taking.
- It promotes good discussion: It does not appear to provide a solution, merely states a need.
- It helps tease out appropriate price points by being written to encourage segmentation on aspects that are important to us: loops versus no loops, time of day, and probably complex road areas that require more data regardless of the application.
- It is legally sound: We follow good procedures and identify the areas that will affect our decisions.

5.3 PURCHASING DATA

There is a natural tension between assessing data quality prior to purchasing and an RFP meant to encourage a broad response. We considered three purchasing possibilities:

- **Option 1**—Buy data per corridor. Quotes are number of points per corridor. No additional information. We would get two quotes (or four quotes if we include direction). No penalty for less data.
- **Option 2**—Buy per important segment: one segment of I- 880 and three segments of I-15. Quotes are prices per segment for number of points per day divided into four quadrants (Morning, Midday, Evening, and Night) and by Weekday (not individual weekdays) and Weekend. We would get four quotes (or eight quotes if we include direction). Respondents are expected to meet their average number 80% of the time averaged over a month or they will only receive 75% per month of their quote.
- **Option 3**—Buy data per corridor but separate the quote granularity from the data quality granularity. Response is scientific formula which requires measuring data per mile per 15 minute intervals and aggregating. Perhaps we would get either two or four quotes this way, but it would require many matrices of data analysis to accompany the quotes.

Each option has pluses and minuses:

- Option 1 is too simple and actually hurts the procurement process by not stimulating discussion, not segmenting, and not providing a way of determining data quality. It is simple, which is a virtue, and does not ask for too much data prior to our buying it. Quote is related to requested metrics.

- Option 2 is a middle ground. It is not simple but not complex either. Provides sufficient data for comparison to others. May promote a discussion. Quote is related to metrics.
- Option 3 is very strong on data quality determination. However, data quality is only one metric we are using in our decision (company, price, and so on). It could be viewed as overly complex by providers who are not data aggregators. Thus, it might reduce the number of responses and narrow the discussion. Also, companies may be quite reluctant to provide the information prior to being chosen. It obscures the segmentation issues by dividing data into statistically significant segments not related to actual data value.

☑ ***What we decided:*** We selected **Option 2** and decided to use the scientific algorithm to evaluate data after we start getting it.

5.4 PRICING

Pricing models for unaggregated probe data sources remain relatively undeveloped. They are inconsistent between vendors and have different structures, depending on the vendor. However, most vendors for aggregated data charge a fairly nominal set-up fee and a more substantial ongoing service fee, typically yearly. The one constant is that nearly all costs are charged in units of directional road miles.

Because of this pricing model immaturity for unaggregated data, we had numerous in-depth discussions on how to ask for the vendor’s pricing model in the RFP and for the payment structure and approach. We considered four options for the cost proposal:

- Flat rate cost per corridor for the duration of the contract
- Flat cost per mile
- Cost per point per mile
- Cost per highway segment for the duration of the contract

☑ ***What we decided:*** We settled on the last option, the estimated price defined as a total charge for each highway segment over the entire duration of the contract (≈ 3 months).

6 SELECTION CRITERIA

In exploring the feasibility of purchasing unaggregated traffic data, we were venturing into unfamiliar territory, even entertaining the idea that such a market might not be available. We therefore had extensive discussions about what criteria could be used to evaluate a vendor's proposal.

The following criteria were considered, although not exclusively, in determining which firm was selected.

6.1 MANAGEMENT (TOTAL 30 POINTS)

- History of work in data architecture, data management, and systems integration
- Personnel with relevant experience
- Ability to isolate PATH commitment from other parts of the business/organization to insulate work on this RFP from stresses in other parts of the business
- Planning and execution experience: track record in contract administration and successful delivery on time and on budget
- Ability to meet requirements and design processes that clearly match the risk and benefits profile associated with the deliverables of this RFP
- Internal controls and strategies in place to ensure that data requirements are met and to address schedule slippages and resource challenges

6.2 DATA (TOTAL 50 POINTS)

- Self-reported data coverage and quantity as presented in the data worksheet. Note that data was subject to validation by PATH.

6.3 TECHNICAL (TOTAL 20 POINTS)

- Traffic sensor deployment and data collection experience
- Traffic data procurement or validation expertise
- Knowledge of traffic information systems

7 SOLICITING PROPOSALS

On May 9, 2011, PATH released the request for proposals, entitled “Unaggregated Data Procurement.” PATH complied with the purchasing methodology outlined in University of California Business and Finance Bulletin BUS-43, particularly Part 3, §III, which refers to common goods, materials, and services over $100,000 in value. This request for proposals (RFP) was issued by PATH on behalf of the Regents of the University of California, who were represented by the Procurement Services Office of the University of California, Berkeley (the Issuing Office).

The solicitation document was made available to potential bidders by public release on PATH’s website, by Business Services with University of California Berkeley, and by direct emails to several vendors identified as having previously engaged in related work. PATH indicated that it wished to welcome bids from a variety of vendors, including those who were not currently prepared to provide coverage of the entire study area. As multiple awards for this process were anticipated, bids with partial coverage were encouraged.

7.1 VENDOR INQUIRIES

PATH offered potential bidders the opportunity to ask questions about the project prior to submitting proposals. These questions were due by May 23, 2011. Two sets of questions were received, and the anonymized questions and PATH’s responses were publicly released on June 6, 2011, on PATH’s website.

7.2 REVIEWING AND SELECTING VENDORS

The due date for proposals was June 20, 2011. We received proposals from four vendors. (The proposals are included in the Supporting Documents for Task Orders 1 and 2, provided to Caltrans in electronic format along with this report.)

On July 21, 2011, the proposals were reviewed by a team consisting of PATH staff (including principal investigator Alex Bayen and engineers Joe Butler, Ali Mortazavi, Nazy Sobhi), Nick Compin of Caltrans, and Janet Banner of MTC. Alex Bayen and Joe Butler were the only voting members of the proposal selection panel.

Using the guidelines described in section 6, *Selection criteria,* the selection panel determined that two proposals substantially met the technical requirements and that the costs were acceptable. One other proposal technically was acceptable, but their cost was very high. The fourth proposal did not meet the technical requirements and was not considered further.

7.3 NEGOTIATING THE CONTRACT

The successful firms were notified that their proposals had been accepted, and UC Berkeley's Purchasing Office wrote the contracts based on the bids received from the vendors. Each side's legal department then reviewed the contract to finalize its terms.

Negotiations involved back and forth discussions of the contract terms to align the legal, regulatory, business, and technical requirements of both parties and reach agreement on such issues as:

- Allowing PATH perpetual use of the purchased data
- Allowing PATH to mix the data with that from other sources
- PATH confirming it would not reveal a firm's data to its competitors
- Price

The contracts are included in the Supporting Documents for Task Orders 1 and 2, provided to Caltrans in electronic format along with this report.

8 RECOMMENDATIONS

The traffic data reporting market is relatively young and growing. Caltrans should continue to monitor the cost and accuracy of probe data while avoiding any decisions that limit future options. It is likely that the cost and accuracy will reach a point where implementation will be cost-effective. Caltrans should continue to engage stakeholders that would make use of probe data and explore the various uses and potential value of this information. Based on the responsive proposals that we received, we are confident in recommending the following transition steps to Caltrans.

8.1 RECOMMENDATIONS FROM PATH'S TRAFFIC DATA PROCUREMENT

- Procure speed data as a product, through an open solicitation, for selected corridors, for a fixed duration.
- Use our identified guidelines for the data specification, data collection description, and presentation of data and definition of coverage metrics to procure speed data.
- Integrate this data into the data management system.

8.2 RECOMMENDATIONS FROM PREVIOUS TRAFFIC DATA PROJECTS

Previously, in a separate task, we conducted a series of interviews with state DOTs (Georgia, Minnesota, Wisconsin, and I-95 Coalition) that have worked with private traffic data providers (see Chapter 2, *Assessment of Current Practices*). Although only two of the projects went through a standard RFP process, we were able to identify the following recommendations based on overall experiences:

- Consider a Request for Information (RFI) to develop specifications for the RFP**.** The I-95 Coalition used two RFIs to refine their RFP specifications. For the first RFI, they shared their vision for the project and elicited ideas from the vendors. They used the responses to develop specifications for the traffic data for a second RFI, which gave the vendors a chance to comment. The result was that the 5–6 proposals submitted for the RFP were all on target.
- Meet early with the procurement staff**.** Since procuring ongoing data services is somewhat different from either goods (computers, office equipment, etc.) or services (catering, consulting, etc.), the procurement staff may need to be educated on some of the differences and the purpose of the ongoing service.
- Include a clear exit strategy. Be explicit about how you can cancel the contract if you are not getting what you want.
- Define optional versus required features. To minimize the chance that no contract is awarded, a set of required features should be distinguished from a set of optional or desired provisions.

This will allow the vendors to have some flexibility in how they respond to the RFP. An alternative way of implementing this would be to have levels of compliance with feature requests—the higher the level of compliance the more points awarded.

- Establish your validation process in advance of the data procurement. Some of the project managers we spoke with mentioned that it was time-consuming to develop the validation process after the data arrived. However, other groups, such as the I-95 Coalition, have continually refined their validation process.

- Create benchmarks for delivering the data feeds. Several of the previous projects experienced delay in receiving the data feeds. Incorporating timely benchmarks into the RFP will force vendors to account for them in their proposal. Vendors should also be asked to describe their current coverage areas and the rate at which they plan to expand geographic coverage and to what areas.

- Create milestones for different phases in the project**.** In addition to timely data delivery, some DOTs specified milestones for acceptance testing and integration into existing DOT systems.

- Consider using a testing stage before moving to contract. Wisconsin DOT did not proceed to a contract with Airsage after the data performed poorly during acceptance testing. This saved the DOT the costs and time of proceeding with a contract that would not bring value.

- Consider an ongoing monitoring system. Although acceptance and validation testing are essential, some DOTs noticed that traffic data providers would sometimes revert to historical or sensor data. A monitoring system could help track this. Several groups mentioned requiring in the RFP an indicator of how real-time the data is. Having this indicator allows DOTs to make better decisions about how they want to use the data. However, changes in the business offerings of traffic data providers may limit this need. For example, INRIX now offers a real-time data feed as well as a data feed that includes historical and other types of data that are not real-time. To the degree that the market has segmented customer demand for real-time versus modeled data, the need for ongoing monitoring is reduced.

- Specify customer service response times. Only Wisconsin DOT required customer service provisions, i.e., high priority problems being responded to within two hours. Some of these may be part of the standard contract offered by private traffic data providers, but this should be confirmed.

9 GLOSSARY

The following terms denote traffic data at successive stages from its collection to its use, as well as other quantities of interest. The definitions below are introduced for convenience and to avoid any confusion between the parties involved, and as such they should be used as transparent structuring components of the discussion.

- **Data feed**: Collection of raw and/or processed data in a usable format
  Ex: PeMS feed of loop detector data, Mobile Millennium segment speed estimates
- **Data source**: Sensing device relying on a well-identified physical phenomenon
  Ex: Loop detectors using magnetic induction, cell phones using GPS triangulation, speed radar using Doppler shift
- **Data type**: Traffic state and its level of aggregation in space and time
  Ex: Point counts over 30 seconds, instantaneous point speed, 1-mile segment speed over 5 minutes
- **Ground truth**: True value of a data type
  Ex: 30-second occupancy at a given location, vehicular travel time over a fixed route
- **Processed data**: Raw data having been subject to processing
  Ex: Counts from loop detectors, segment speeds from a cell phone using GPS triangulation
- **Processing**: Irreversible transformation
  Ex: Averaging, outliers removal, low-pass filtering
- **Raw data**: Output stream of a data source
  Ex: (Time, device identity, position) from a cell phone using GPS triangulation, (time, device identity, voltage) from a loop detector
- **Reference state**: Best available estimate of ground truth
  Ex: Camera record for vehicle position, Mobile Millennium output for segment speeds
- **Traffic application**: Activity of interest for a transportation organization
  Ex: Point-speed estimation, ramp metering, variable toll rates
- **Traffic state**: Fundamental traffic quantity
  Ex: Count, occupancy, density, speed, vehicle miles traveled

Chapter 5

Data Fusion

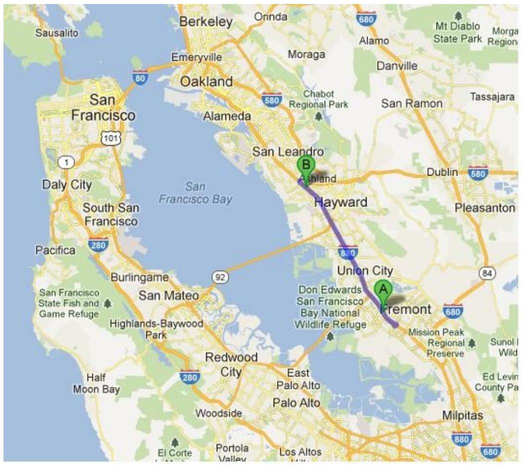
After the contracts for purchasing probe data were executed and the data feeds set up, the research team began working with the procured data to see if it could be used, evaluated, and fused with loop detector data to estimate travel times along the highway. This chapter details the steps taken and the results achieved in pursuing those objectives.

1 EXPERIMENTAL CONTEXT

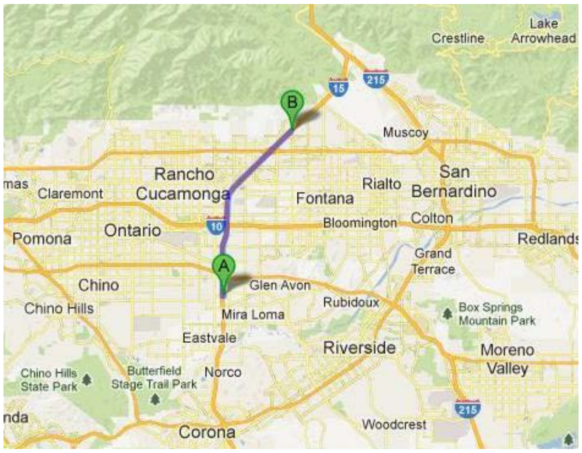
1.1 THE ENVIRONMENT

As described in Chapter 3, we studied three segments of highway, one in the Bay Area and two in southern California:

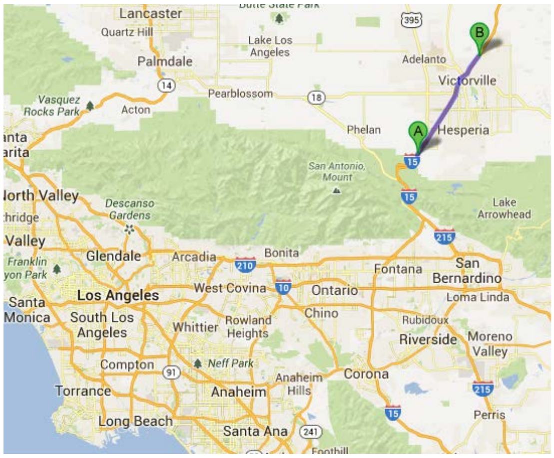
- I-880 from Fremont to Castro Valley
- I-15 in Ontario
- I-15 in Victorville



I-880



I-15 Ontario



I-15 Victorville

This chapter provides a detailed description of our work on I-880 northbound. Similar results were achieved for I-15 and are summarized later in this chapter.

1.2 THE DATA FLOW PROCESS

Figure 5-1 illustrates the data flow process for this study:

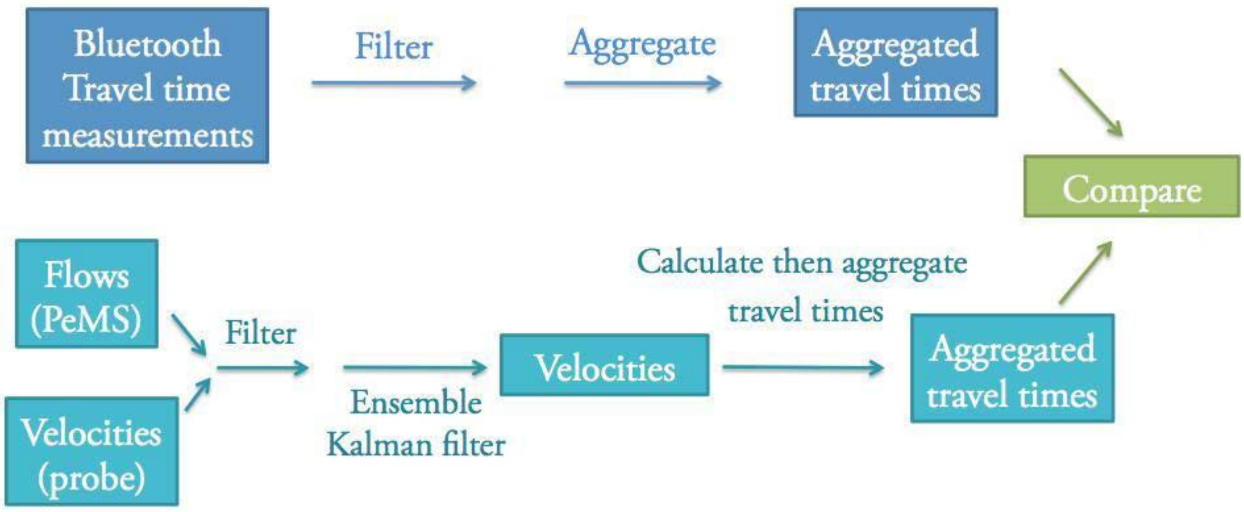


Figure 5-1: Diagram of the data flow process

Bluetooth measurements. As detailed in Chapter 3, the information obtained from Bluetooth detectors is the travel times of vehicles on each of the nine sections of highway bounded by the ten Bluetooth detectors that we deployed along I-880. Since the Bluetooth sensors detect individual devices, that information must be filtered and aggregated: We must remove duplicate and outlying data points and take the average of the travel times measured over a given period of time to be able to compare it with other sources of data.

PeMS data. With loop detector data from PeMS, we measure vehicle counts and occupancies at specific locations along the roadway, and apply a filter to remove faulty data and obtain densities. We then use the fundamental diagram (described in section 2.2) to calculate velocities.

Probe data. With probe data, we get a list of velocities associated with positions of vehicles equipped with GPS devices. Probe data is run through the *path inference filter* (PIF) to accurately map the probe vehicles to the road network and plot their trajectories. The filter map-matches each probe data point to one or several possible nearby points in the network, associated with probabilities. We decided to retain all the map-matched data points with probability greater than 0.7. Thus, only trusted data (with a probability 0.7) is kept and used in the assimilation. (The path inference filter is described in the final report for Task Order 2, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods*.)

Travel time estimation. In this study, we use a velocity estimation algorithm developed at Berkeley as part of the Mobile Millennium project. The algorithm (or "data fusion engine") combines velocity measurements from probe vehicles or inductive loop detectors with a model of traffic evolution, using a technique known as *ensemble Kalman filtering* (EnKF), to produce an improved estimate of the velocity field, from which the travel time is computed. The resulting travel times computed from this process are then compared to the travel times computed from the Bluetooth measurements.

Probe data volume and storage. There was a stark contrast in the amount of probe data we received from the two vendors who were selected for the project. One vendor delivered data for the entire U.S., averaging 720,000,000 data points per day. After narrowing this down to the areas of the study, we retained an average of 36,000,000 data points per day. The other vendor supplied an average of 180,000 data points per day, focused on the sites of interest. For the two months of the study, we found that the data could be stored effectively with one terabyte (1TB) of disk space.

2 MATHEMATICAL CONTEXT: MODEL OF TRAFFIC EVOLUTION

2.1 THE LWR EQUATION

The evolution of the density of vehicles is described by a conservation law which states that the variation of the number of vehicles in a cell of infinitesimal length dx during an infinitesimal time dt is equal to the difference between the number of vehicles entering the cell and the number of vehicles exiting the cell during dt. The whole macroscopic continuous flow model brings us to the equation known as the *Lighthill-Whitham-Richards* (LWR) *partial differential equation* (PDE), and describes the evolution of vehicle density ρ for a stretch of highway from point a to point b over a time T.

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(\rho(x,t))}{\partial x} = 0 \quad (x,t) \in (0,L) \times (0,T) \tag{1}$$

where the flux function $Q(\cdot)$ expresses the flow of vehicles as a function of the density, and is known as the *fundamental diagram* in the transportation engineering community.

Another way to write Equation (1) using the chain rule is:

$$\frac{\partial \rho(x,t)}{\partial t} + Q'(\rho(x,t))\frac{\partial \rho(x,t)}{\partial x} = 0 \tag{2}$$

To use this model on a bounded portion of highway, it is necessary to specify boundary and initial conditions:

$$\begin{cases} \rho(x,0) = \rho_0(x) \forall x \in [a,b] \\ \rho(a,t) = \rho_a(t), \rho(b,t) = \rho_b(t) \forall t \in [0,T] \end{cases} \tag{3}$$

Assuming that the velocity can be modeled as a function $V(\cdot)$ of the density, the flux function reads:

$$Q(\rho) = \rho V(\rho) \tag{4}$$

2.2 THE FUNDAMENTAL DIAGRAM

From experimental observations, it appears that there exists a correlation between speed, flow, and density. In other words, the functions $q = Q(\rho)$ and $v = V(\rho)$ exist. The first to mention this was Greenshields in 1934 in [1]. The relation $q = Q(\rho)$ is commonly called *the fundamental diagram*. Figure 5-2 shows a sample of those most commonly used:

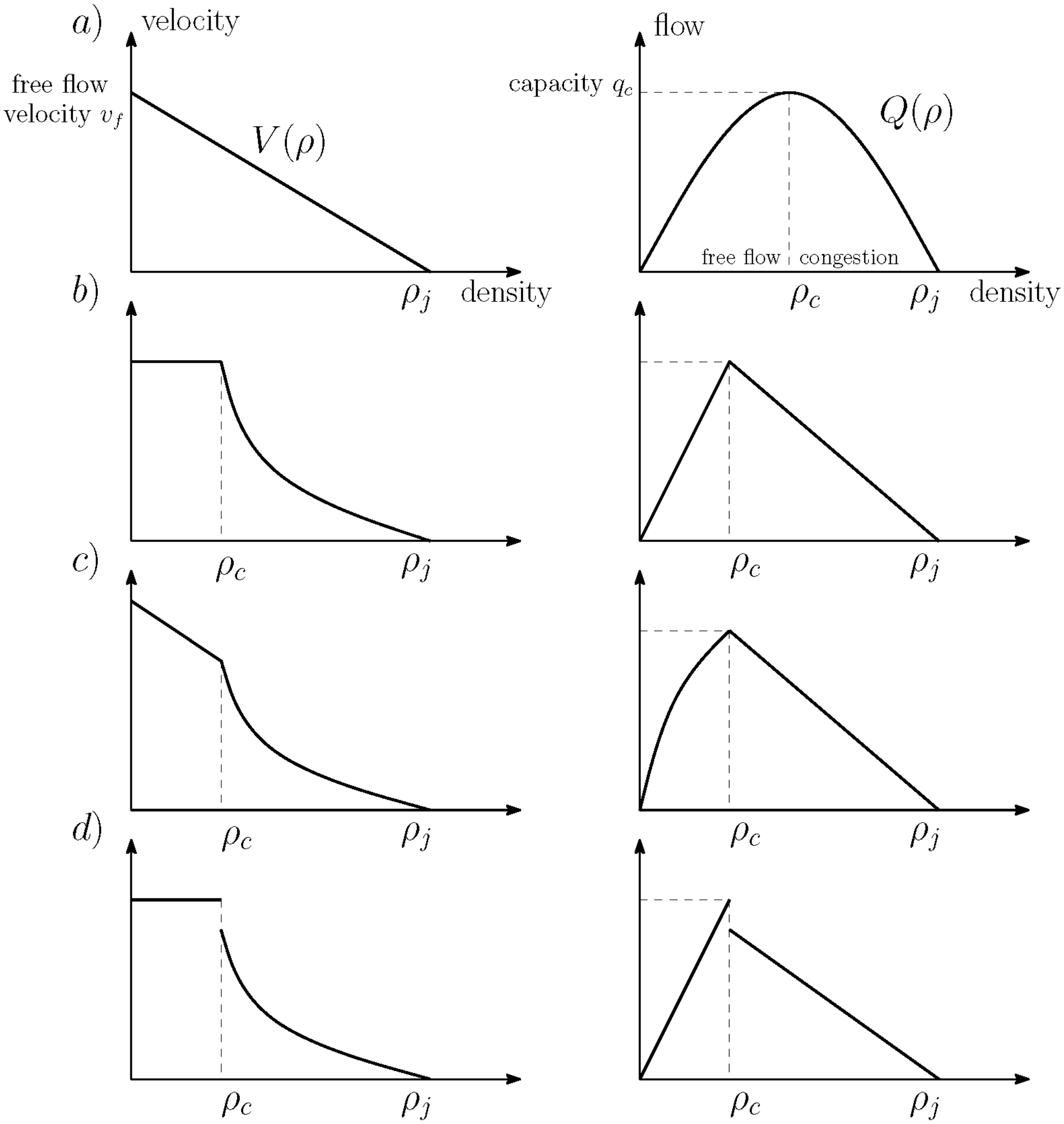


Figure 5-2: Speed and flow as a function of density for Greenshields (a), Daganzo-Newell (b), linear-hyperbolic (c), and discontinuous (d) fundamental diagrams. In this study, we use the linear-hyperbolic fundamental diagram (c).

The Greenshields flow is simple and convenient to use because of the regularity of the functions, but experimental studies have shown that the Daganzo-Newell is one of the most realistic and accurate models which verifies concavity properties.

Greenshields fundamental diagram:

$$v = V_G(\rho) = v_f(1 - \frac{\rho}{\rho_j}) \tag{5}$$

and the corresponding flux function is:

$$Q_G(\rho) = \rho V_G(\rho) = v_f(\rho - \frac{\rho^2}{\rho_j}) \tag{6}$$

Daganzo-Newell fundamental diagram:

$$v = V_{DN}(\rho) = \begin{cases} v_f & if \quad \rho \leq \rho_c \\ -w\left(1 - \frac{\rho_j}{\rho}\right) & if \quad \rho > \rho_c \end{cases} \tag{7}$$

and the corresponding flux function is:

$$Q_{DN}(\rho) = \rho V_{DN}(\rho) \begin{cases} v_f \rho & if \quad \rho \leq \rho_c \\ -w(\rho - \rho_j) & if \quad \rho > \rho_c \end{cases} \tag{8}$$

where v_f is the *free-flow speed*, ρ_j is the *jam density* (i.e., $Q(\rho_j) = 0$), ρ_c is the *critical density* (i.e. *free flow* when $\rho \leq \rho_c$ and *congestion* when $\rho > \rho_c$) and w is the backward wave propagation speed.

It is important to notice that both of the previous fundamental diagrams satisfy the following properties:

(i) Hypothesis of a static flow/density relationship: $q = Q(\rho(x,t))$.

(ii) $Q(0) = Q(\rho_j) = 0$.

(iii) The continuous portions of $Q(\rho)$ are concave.

(iv) $V(0) = v_f$, and $V(\rho_j) = 0$.

(v) A critical density ρ_c can be defined where the maximum flow q_c is attained. Then, $Q(\rho)$ is increasing for $\rho \leq \rho_c$ and decreasing for $\rho > \rho_c$.

(vi) The critical density ρ_c separates the fundamental diagram into two regimes: *free flow* when $\rho \leq \rho_c$ and *congestion* when $\rho > \rho_c$.

The Daganzo-Newell is the most commonly used, but because its velocity function is constant in free-flow, it cannot be inverted. This is why in our model we use the linear-hyperbolic fundamental diagram, which is close to the Daganzo-Newell and has the advantage of having an invertible velocity function. This is necessary because the model uses velocity measurements and not densities. Thus, all the results presented in this chapter have been obtained using this flux function:

$$Q(\rho) = \rho V(\rho) \tag{9}$$

which becomes

$$\tilde{Q}(v) = V^{-1}(v)v \tag{10}$$

and we define the linear-hyperbolic flow function with a linear expression in free-flow and a hyperbolic expression in congestion.

Linear-hyperbolic fundamental diagram:

$$v = V_{LH}(\rho) = \begin{cases} v_f\left(1 - \dfrac{\rho}{\rho_j}\right) & if \quad \rho \leq \rho_c \\ -w\left(1 - \dfrac{\rho_j}{\rho}\right) & if \quad \rho > \rho_c \end{cases} \tag{11}$$

Its inverse :

$$\rho = V_{LH}^{-1}(v) = \begin{cases} \rho_j\left(1 - \dfrac{v}{v_f}\right) & if \quad v \geq v_c \\ \rho_j\left(\dfrac{w}{w+v}\right) & if \quad v < v_c \end{cases} \tag{12}$$

where v_c is the critical velocity $v_c = V(\rho_c)$ and where the relation $\frac{\rho_c}{\rho_j} = \frac{w}{v_f}$ must be satisfied to ensure continuity of the flux at the critical density ρ_c.

The corresponding flux function is:

$$Q_{LH}(\rho) = \rho V_{LH}(\rho) \begin{cases} v_f\left(\rho - \dfrac{\rho^2}{\rho_j}\right) & if \quad \rho \leq \rho_c \\ -w(\rho - \rho_j) & if \quad \rho > \rho_c \end{cases} \tag{13}$$

and in terms of velocities, the flux function becomes, using 10 :

$$\tilde{Q}_{LH}(v) = \begin{cases} \rho_j\left(v - \dfrac{v^2}{v_f}\right) & if \quad v \geq v_c \\ \rho_j w\left(1 - \dfrac{w}{w+v}\right) & if \quad v < v_c \end{cases} \tag{14}$$

While measurements in the free-flow part are usually represented by a continuous curve, the congestion measurements are often more scattered. Some authors argue that this fact can be modeled by introducing a discontinuity at $\rho = \rho_c$ as in Figure 5-2 part (d).

2.3 THE CELL TRANSMISSION MODEL (CTM)

To solve the partial differential equations along the road network, we follow the classical path using the Godunov scheme, also called Cell Transmission Model. This numerical scheme is based on exact solutions to Riemann problems [2] and leads to the construction of a nonlinear discrete-time dynamical system.

2.3.1 THE GODUNOV SCHEME

The Godunov discretization scheme is applied to the LWR PDE, where the discrete five minutes per mile ΔT is indexed by n, and the discrete space step ΔX is indexed by i:

$$\rho_i^{n+1} = \rho_i^n - \frac{\Delta T}{\Delta X}\left(G(\rho_i^n, \rho_{i+1}^n) - G(\rho_{i-1}^n, \rho_i^n)\right) \tag{15}$$

where the Godunov flux $G(\rho_1, \rho_2)$ is in general defined as:

$$G(\rho_1, \rho_2) = \begin{cases} \min_{\rho\in[\rho_1,\rho_2]} Q(\rho) & \text{if} \quad \rho_1 \le \rho_2 \\ \max_{\rho\in[\rho_2,\rho_1]} Q(\rho) & \text{if} \quad \rho_2 \le \rho_1 \end{cases} \tag{16}$$

Similarly, we can apply the scheme at the boundaries (see [3])

$$\rho_0^{n+1} = \rho_0^n - \frac{\Delta T}{\Delta X}\left(G(\rho_0^n, \rho_1^n) - G(\rho_{-1}^n, \rho_0^n)\right) \tag{17}$$

$$\rho_{i_{max}}^{n+1} = \rho_{i_{max}}^n - \frac{\Delta T}{\Delta X}\left(G(\rho_{i_{max}}^n, \rho_{i_{max}+1}^n) - G(\rho_{i_{max}-1}^n, \rho_{i_{max}}^n)\right) \tag{18}$$

with the ghost cells on the left and the right of the space interval defined as follows:

$$\rho_{-1}^n = \frac{1}{\Delta t}\int_{(n-\frac{1}{2})\Delta t}^{(n+\frac{1}{2})\Delta t} \rho_a(t)\,\mathrm{d}t. \tag{19}$$

$$\rho_{i_{max}+1}^n = \frac{1}{\Delta t}\int_{(n-\frac{1}{2})\Delta t}^{(n+\frac{1}{2})\Delta t} \rho_b(t)\,\mathrm{d}t. \tag{20}$$

2.3.2 THE CFL CONDITION

In order to ensure numerical stability, the time and space steps are coupled by the CFL condition. In the case of our study which is a hyperbolic PDE, the CFL condition becomes:

$$\frac{\Delta T}{\Delta X} \leq \frac{1}{c_{max}} \tag{21}$$

where c_{max} denotes the maximal characteristic speed.

In our case, $\Delta T = 6$ seconds and ΔX takes different values depending on the location on the route, but in every case, the CFL condition is verified.

2.3.3 THE V-CTM

In practice, probe data directly measure velocities, and loop detectors measure densities which are converted into velocities using the fundamental diagram. The v-CTM is obtained by modifying the Godunov scheme to use velocities instead of densities.

Using the definition of $\tilde{Q}$ in (10), we can transform (15) to obtain :

$$v_i^{n+1} = V\left(V^{-1}(v_i^n - \frac{\Delta T}{\Delta X}\left(\tilde{G}(v_i^n, v_{i+1}^n) - \tilde{G}(v_{i-1}^n, v_i^n)\right)\right) \tag{22}$$

and the Godunov flux $\tilde{G}(v_1, v_2) = G(V^{-1}(v_1), V^{-1}(v_2))$ becomes:

$$\tilde{G}(v_1, v_2) = \begin{cases} \max_{v \in [v_1, v_2]} \tilde{Q}(v) & \text{if} \quad v_1 \leq v_2 \\ \min_{v \in [v_2, v_1]} \tilde{Q}(v) & \text{if} \quad v_2 \leq v_1 \end{cases} \tag{23}$$

2.4 DATA ASSIMILATION

2.4.1 THE KALMAN FILTER

Using the velocity measurements of the probe or loop detector data, the idea is now to estimate the velocity v^n over the whole network at each five minutes per mile. Given the velocity field on the network at time $n\Delta t$, the velocity at time $(n+1)\Delta T$ is constructed using the v-CTM algorithm $v^{n+1} = \mathcal{M}[v^n]$ of section 2.3.3. For estimation purposes, we extend the model to

$$v^n = \mathcal{M}[v^{n-1}] + \eta^n \tag{24}$$

where $\eta^n \sim (0, \boldsymbol{Q}^n)$ is the Gaussian zero-mean, white state noise with covariance $\boldsymbol{Q}^n$, used to model inaccuracies in the evolution model.

Each velocity measurement from mobile devices can be modeled with a linear observation operator given by :

$$y^n = \boldsymbol{H}^n v^n + \chi^n \tag{25}$$

where the linear observation matrix $\boldsymbol{H}^n \in \{0,1\}^{p^n \times \kappa}$ encodes the p^n discrete cells on the highway for which the velocity is observed during discrete five minutes per mile n and κ is the corresponding (total) number of cells in the network. The last term in expression (25) is the white, zero-mean observation noise $\chi^n \sim (0, \boldsymbol{R}^n)$ with covariance matrix $\boldsymbol{R}^n$.

Based on [4] which gives a general formulation of the Kalman Filter for discrete time systems, [5] gives a formulation of the Ensemble Kalman filter for the velocity-based model. The simple Kalman filter cannot be used because of the nonlinearity of the model. If equation (22) was differentiable in v^n, the operator $\mathcal{M}[\cdot]$ in (24) would be as well, and we could have obtained the estimate v^n with the extended Kalman filter for nonlinear systems. But here $\mathcal{M}[\cdot]$ is not differentiable, which is why we use the Ensemble Kalman filter.

2.4.2 THE ENSEMBLE KALMAN FILTER

The ensemble Kalman filter was introduced by Evensen in [4] to overcome the difficulties due to nonlinear state evolution models in the EKF. By using ensemble integrations and propagating the ensemble of model states forward in time, the EnKF makes it possible to calculate the mean and the covariances of the error needed at the analysis (measurement-update) step [6]. Two other advantages of the EnKF are that it uses the standard update equations of EKF (except that the gain is computed from the error covariances provided by the ensemble of model states) and that it comes with a relatively low numerical cost. Indeed, only a rather limited number of ensemble members are usually needed to achieve a reasonable statistical convergence [6]. In our case for example, only 100 ensembles are used.

In the traditional Kalman filter, the error covariance matrices are defined as $\boldsymbol{P}_f = E[(v_f - v_t)(v_f - v_t)^T]$ and $\boldsymbol{P}_a = E[(v_a - v_t)(v_a - v_t)^T]$ where $E[\cdot]$ is the average over the ensemble, v is the model state vector at particular time, and the subscripts f, a, and t represent the forecast, analyzed, and true state, respectively. In the EnKF, because the true state is not known, ensemble covariances have to be considered and the matrices are evaluated around the ensemble mean $\bar{v}$, yielding $\boldsymbol{P}_f \approx \boldsymbol{P}_{\text{ens},f} = E[(v_f - \bar{v}_f)(v_f - \bar{v}_f))^T]$ and $\boldsymbol{P}_a \approx \boldsymbol{P}_{\text{ens},a} = E[(v_a - \bar{v}_a)(v_a - \bar{v}_a)^T]$ where the subscript ens refers to the ensemble approximation. In [6], it is shown that if the ensemble mean is used as the best estimate, the ensemble covariance can be interpreted as the error covariance of the best estimate. The ensemble Kalman filter algorithm can be summarized as follows [4]:

1. *Initialization*: Draw K ensemble realizations $v_a^0(k)$ (with $k \in \{1, \cdots, K\}$) from a process with a mean speed $\bar{v}_a^0$ and covariance $\boldsymbol{P}_a^0$.
2. *Forecast*: Update each of the K ensemble members according to the v-CTM forward simulation algorithm. Then update the ensemble mean and covariance according to:

$$v_f^n(k) = \mathcal{M}[v_a^{n-1}(k)] + \eta^n(k) \tag{26}$$

$$\bar{v}_f^n = \frac{1}{K}\sum_{k=1}^{K} v_f^n(k) \tag{27}$$

$$\boldsymbol{P}_{\text{ens},f}^n = \frac{1}{K-1}\sum_{k=1}^{K}\left(v_f^n(k) - \bar{v}_f^n\right)\left(v_f^n(k) - \bar{v}_f^n\right)^T \tag{28}$$

3. *Analysis*: Obtain measurements, compute the Kalman gain, and update the network forecast:

$$\boldsymbol{G}_{\text{ens}}^n = \boldsymbol{P}_{\text{ens},f}^n(\boldsymbol{H}^n)^T\left(\boldsymbol{H}^n\boldsymbol{P}_{\text{ens},f}^n(\boldsymbol{H}^n)^T + \boldsymbol{R}^n\right)^{-1} \tag{29}$$

$$v_a^n(k) = v_f^n(k) + \boldsymbol{G}_{\text{ens}}^n\left(y_{\text{meas}}^n - \boldsymbol{H}^n v_f^n(k) + \chi^n(k)\right) \tag{30}$$

4. Return to 2.

In (30), an important step is that at measurement times, each measurement is represented by an ensemble. This ensemble has the actual measurement as the mean and the variance of the ensemble is used to represent the measurement errors. The perturbations $\chi^n(k)$, which are added to the measurements, are drawn from a distribution with zero mean and covariance equal to the measurement error covariance matrix $\boldsymbol{R}^n$.

In our particular case, the forecast step takes place every six seconds and the analysis step every thirty seconds, every five forecast steps.

3 USING BLUETOOTH DATA

This section presents Bluetooth data that was used to measure ground truth travel times. As described in Chapter 3, Bluetooth detectors were placed along three segments of highway (I-880, I-15 Ontario, and I-15 Victorville) for a period of two weeks at each site. Bluetooth MAC addresses reidentified at different Bluetooth stations provided travel times that were then aggregated over each spatial sub-segment of the road and each time period.

To illustrate our research process, this chapter details the analysis performed on I-880 northbound (unless stated otherwise). Similar results were achieved for I-15 and are summarized later in the chapter.

From March 2 to March 17, 2012, ten Bluetooth detectors (represented by the blue pins) were installed on the portion of highway I-880 shown here, separating the total route (about twelve miles long) into nine smaller ones:

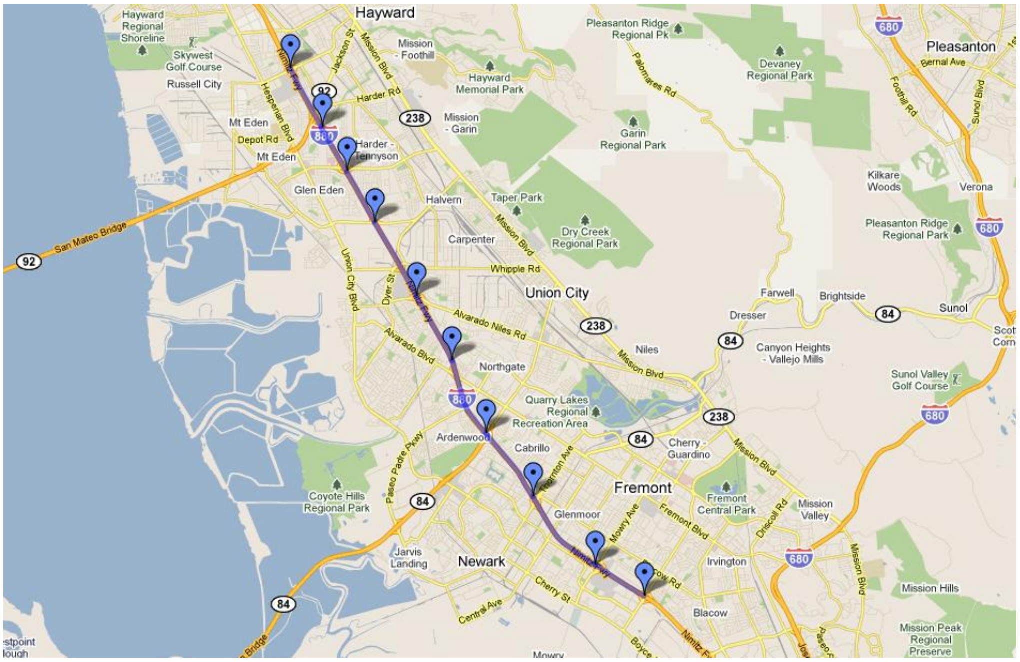


Figure 5-3: The portion of highway I-880 studied

3.1 BLUETOOTH TRAVEL TIME OVER TWO WEEKS

Travel times are grouped into rectangular regions of space-time as depicted in Figure 5-4. The mean and standard deviation of travel times are calculated for each region. Since trajectories may not begin and end in the same region, the beginning of the trajectory is used for aggregation purposes. During the day, there are on average 60 unique detections per 15-minute interval.

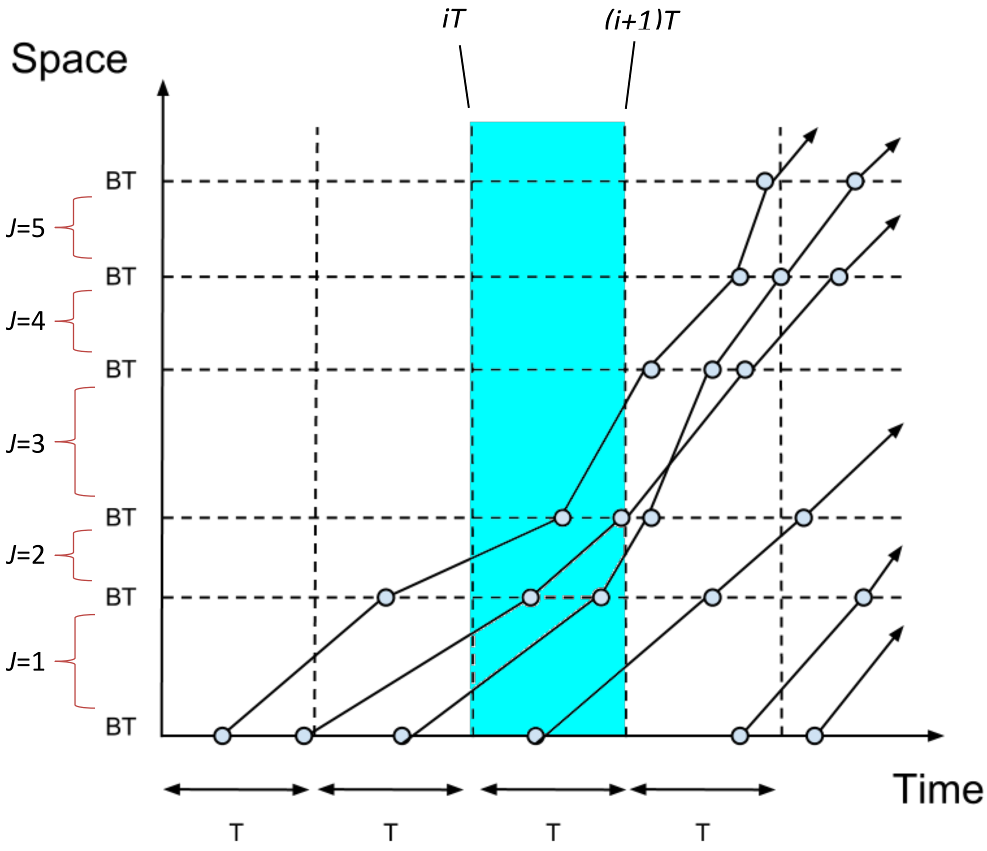


Figure 5-4: Schematic representation of vehicle trajectories along a study site. Vehicles move from bottom to top. Bluetooth detectors are labeled on the y-axis and distributed unevenly along the road. Circles represent detections of a vehicle. Spatial segments (also called routes) between successive Bluetooth devices are indexed by j. The time axis is quantized into segments of duration T and indexed by i.

Individual vehicular travel times are measured by Bluetooth detectors and denoted τ_{jk}, where k is the start time of the vehicle at the beginning of route j. At each time-space region indexed by (i,j) average and standard deviations of travel times are calculated:

$$T_{bt}(i,j) = \frac{1}{\sum_{iT \le k < (i+1)T} 1} \sum_{iT \le k < (i+1)T} \tau_{jk}$$

$$S_{bt}(i,j) = \sqrt{\frac{1}{\sum_{iT \le k < (i+1)T} 1} \sum_{iT \le k < (i+1)T} (\tau_{jk} - T_{bt}(i,j))^2}$$

Two weeks of Bluetooth travel time data is summarized in Figure 5-5 and Figure 5-6. Note that:

- Regions of congestion are numbered and identified within black bounding boxes in Figure 5-5 and Figure 5-6.
- Metrics that are referred to as “global” in this chapter are calculated with respect to those numbered regions only.

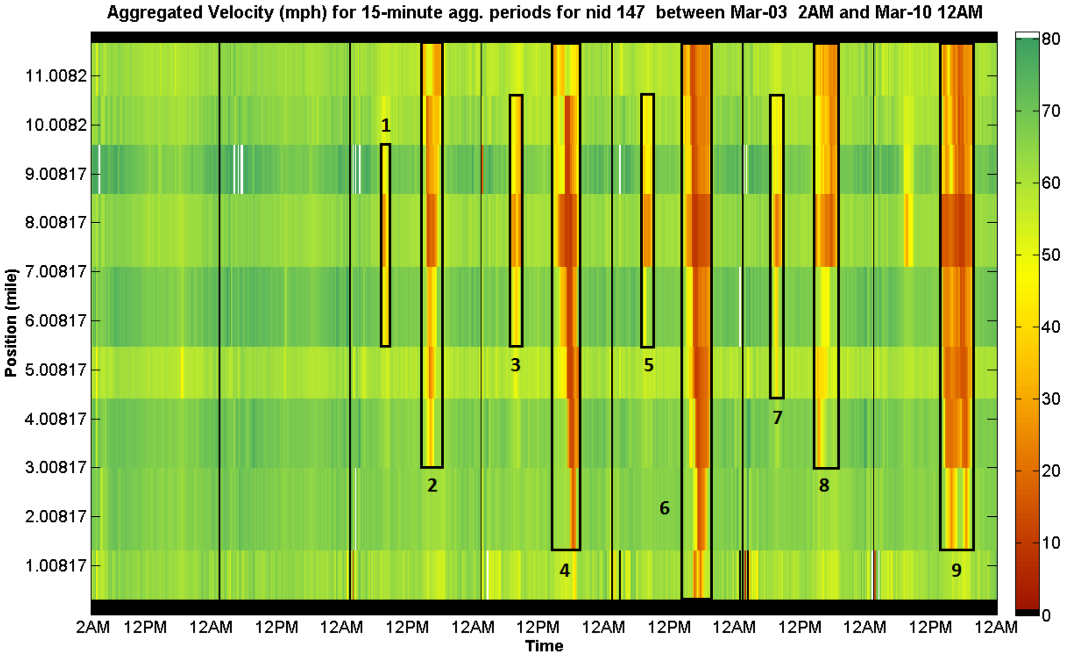


Figure 5-5: Ground truth traffic conditions as derived from Bluetooth travel times and presented as average velocities. (Scale on the right indicates miles per hour.) The data is for one week from Saturday, March 3 to Friday, March 9.

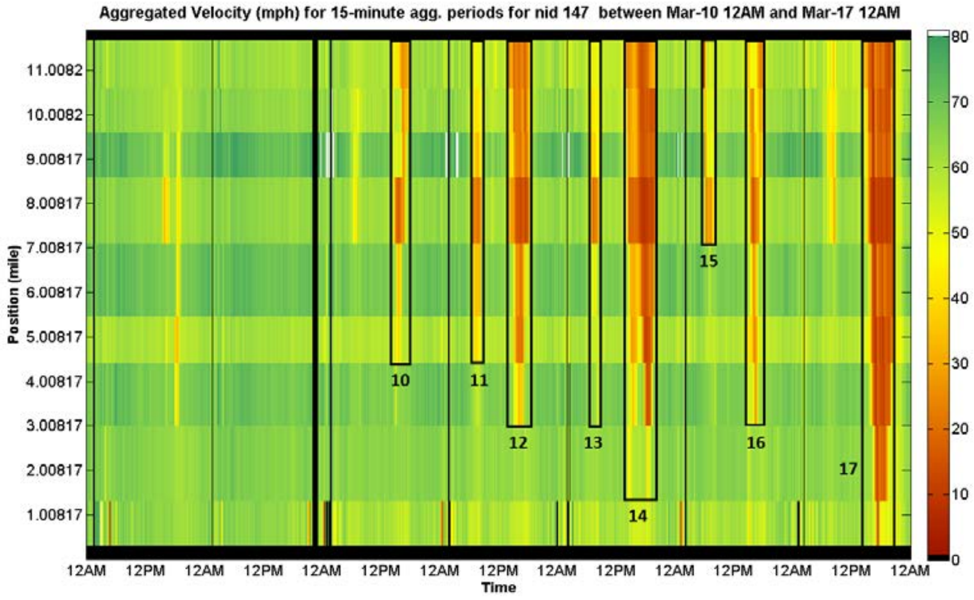


Figure 5-6: Ground truth traffic conditions as derived from Bluetooth travel times and presented as average velocities. The data is for one week from Saturday, March 10 to Friday, March 16.

3.2 VARIATION OF BLUETOOTH DATA OVER ONE DAY

Variability of measured travel times. Even after removing outliers, as explained in Chapter 3, there remains inherent variability in measured travel times. This variability can result from differences in individual driver behavior or from situational differences among multiple lanes of travel.

Instantaneous travel time calculation. For this study, the calculation of travel times uses the notion of “instantaneous” travel times. For this calculation, travel time means and variances are summed over the vertical columns of space-time regions shown in Figure 5-4.

$$T_{bt}(i) = \sum_{j} T_{bt}(i,j)$$

$$S_{bt}(i) = \sqrt{\sum_{j} S_{bt}^{2}(i,j)}$$

In this case, the mean travel time, $T_{bt}(i)$, is a combination of one set of vehicles moving from point A to B, a second (overlapping) set of vehicles moving from B to C, and subsequent sets of vehicles (9 routes). The result of this calculation is displayed in Figure 5-7.

We calculate the variance of the total instantaneous travel time under the assumption of the independence of the partial travel times. This assumption is obviously erroneous but greatly simplifies the calculation. One must keep in mind that the real standard deviation will be greater than the one expressed in the previous formula and in Figure 5-7.

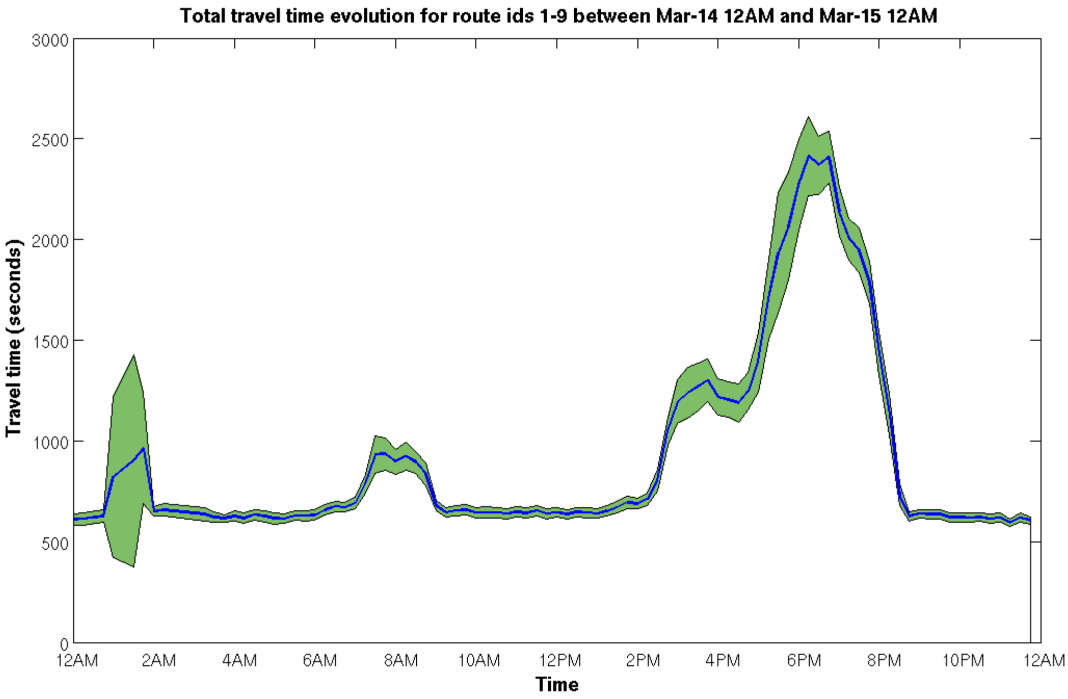


Figure 5-7: Range of travel time measurements plotted against start time. The mean is shown in blue. The green range indicates one standard deviation above or below the mean.

Variability in congestion. Travel time variability is greater during congestion than it is during free flow conditions. For example, near the peak at 6pm in Figure 5-7, total travel time ranged between 37 and 43 minutes. This variability inherent in measured travel times amounts to about 17 percent. Later in this report, the variability observed in Bluetooth measured travel times is used to generate a notion of a “noise floor” for travel time estimation.

4 OVERVIEW OF DATA ASSIMILATION

This section introduces the data types to be assimilated, showcases high-resolution reconstruction of traffic state, and visually compares the results with the measured Bluetooth travel times.

4.1 INTRODUCTION TO PEMS AND PROBE DATA

Data is sparse and interpolation is necessary. Traffic state must be reconstructed over both space and time. However, loop detectors only provide information about traffic state at fixed locations on the road. Most of the road is not covered by any detector, and so the traffic state between detectors must be estimated somehow. Likewise, probe data only provides information at fixed points in both space and time. In both cases, a method must be used to infer traffic conditions between data points in a way that is consistent with the physics of traffic.

Filtered data used in data assimilation. Raw loop detector data is filtered to remove outliers and faulty data and shown in Figure 5-8 (left). Raw probe data is projected onto the road using the path inference filter (described in the final report for Task Order 2, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods*) and shown in Figure 5-8 (right).

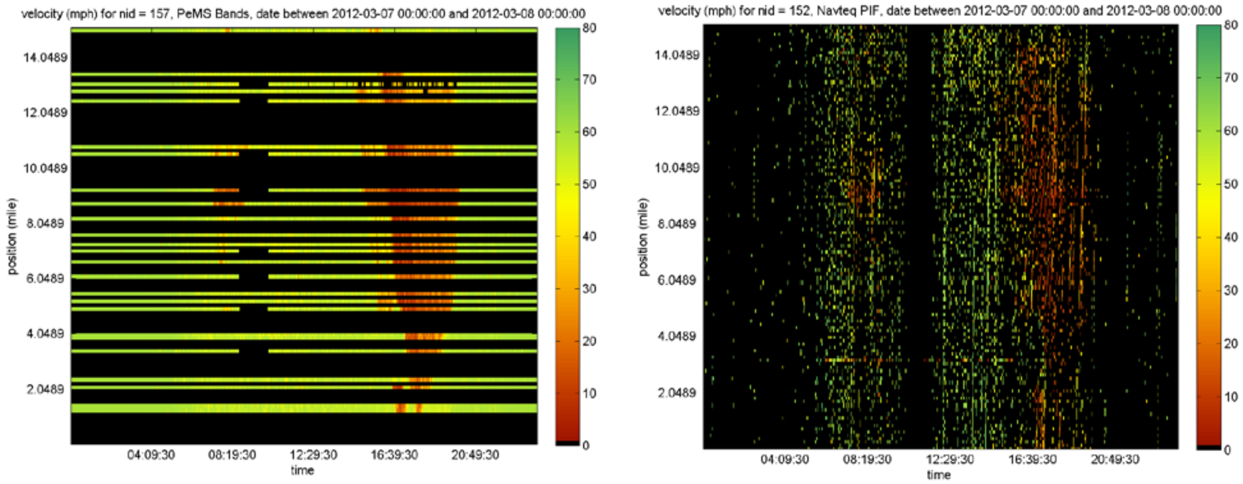


Figure 5-8: Distribution of data along 15 miles of I-880 during March 7 (a typical 24-hour period). Loop data is on the left and probe data is on the right. The color scale between red and green corresponds to slow and fast speeds. Black corresponds to regions of space-time with no data.

4.2 CTM+ENKF VELOCITY MAPS

Visual results of data assimilation. The data fusion engine (the velocity estimation algorithm) takes as input the filtered data from loops and probes and reconstructs the traffic state over the study site, using the well-known CTM and EnKF as described in sections 2.3 and 2.4. The spatial resolution varies slightly depending on road geometry but is approximately 200 meters. The time resolution is 30 seconds.

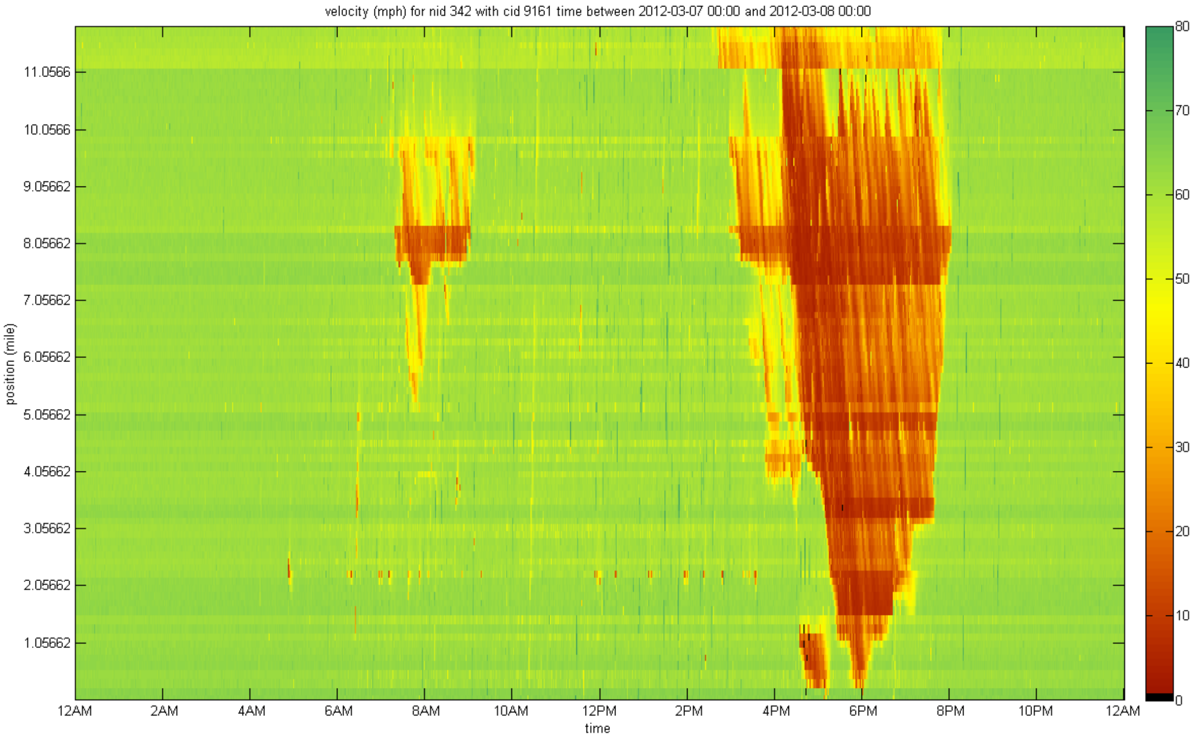


Figure 5-9: Reconstruction of traffic velocity state on March 7 using both loop data and probe data.

Artifacts of input data. Note that fixed loop detector data often appears as faint horizontal streaks, while probe data appears as faint streaks in the direction of vehicle trajectories. We can easily see the backward propagation wave, which agrees with theoretical traffic principles.

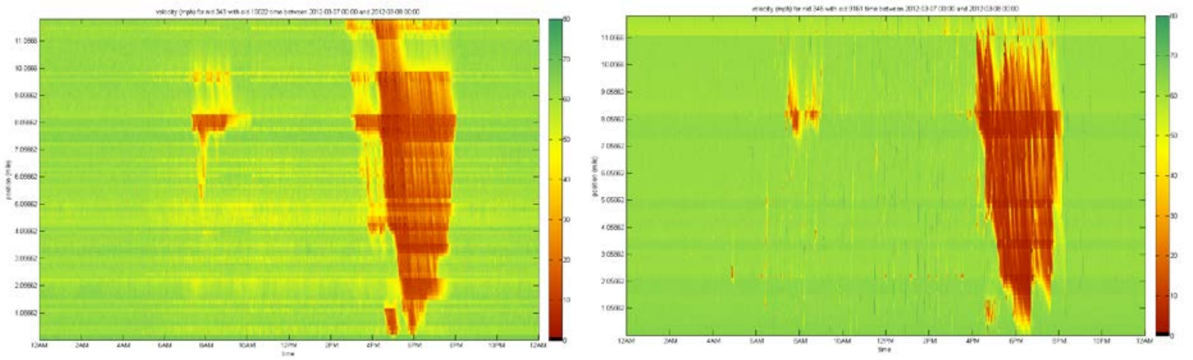


Figure 5-10: Reconstruction of traffic velocity state using loop data only (left) and probe data only (right). These results cover the same regions of space-time along I-880 as shown in Figure 5-9.

4.3 COMPARISON WITH BLUETOOTH AGGREGATED VELOCITY MAPS

Bluetooth used as ground truth. As described in section 3, the quality of the traffic state reconstruction is assessed by comparing it to sampled travel times from Bluetooth detectors. Although presented here as velocities for clarity, the raw measurements being compared are travel times and not velocities.

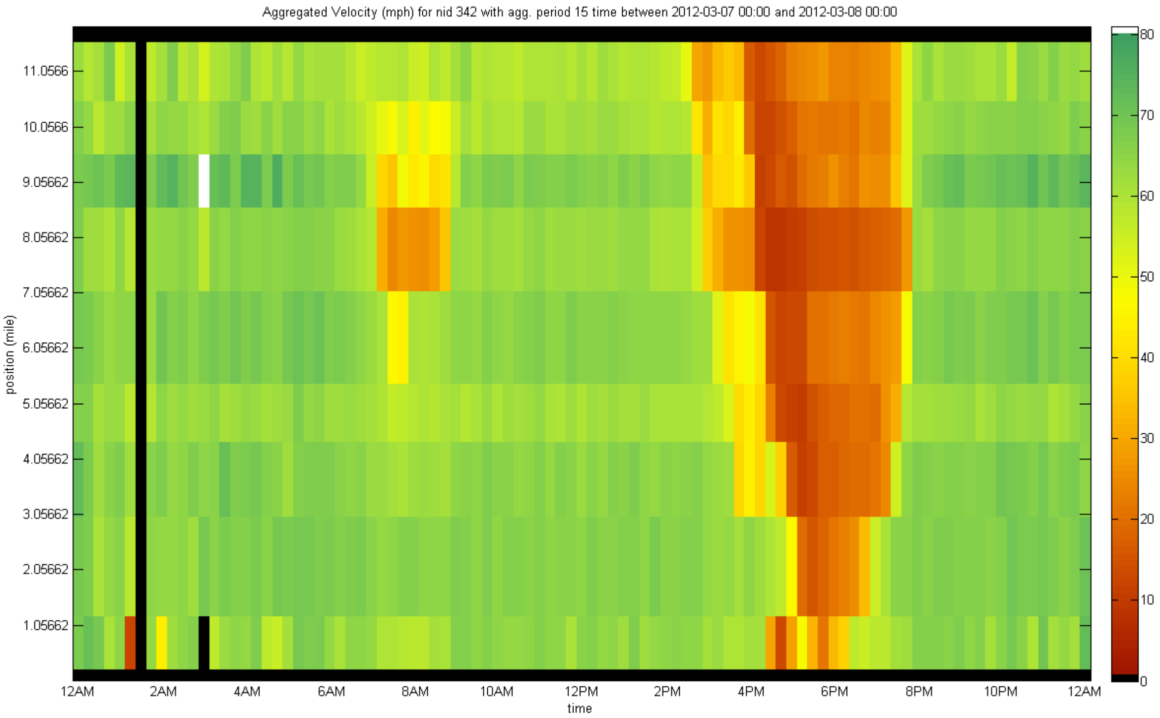


Figure 5-11: Average velocities on March 7 measured over 15-minute intervals along each of nine routes between successive Bluetooth detectors. (Scale on the right indicates miles per hour.)

Walking the velocity map. In order to have an apples-to-apples comparison, the estimated velocities from the data fusion engine (as shown in section 4.2) must be used to generate vehicle trajectories between the same nine routes between successive Bluetooth detectors as shown in Figure 5-11. These generated trajectories are then used to calculate a set of travel times, τ'_{jk}. These travel times are then averaged in a way consistent with the processing of Bluetooth data. Again, time is indexed by i, while j and k designate, respectively, the start route and start time of the vehicles traveling.

$$T_{hwy}(i,j) = \frac{1}{\sum_{iT\le k<(i+1)T} 1} \sum_{iT\le k<(i+1)T} \tau'_{jk}$$

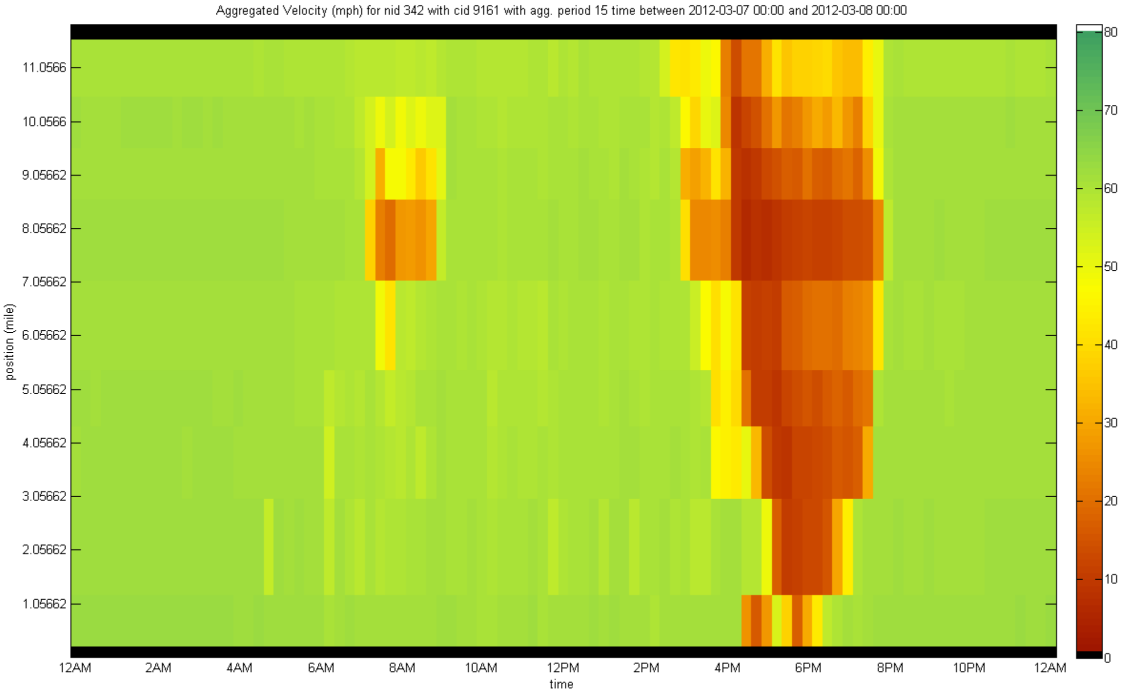


Figure 5-12: Average travel times presented as velocities on March 7 calculated from the reconstruction of vehicle trajectories using both loop and probe data. The travel times are calculated over 15-minute intervals along each of nine routes between successive Bluetooth detectors.

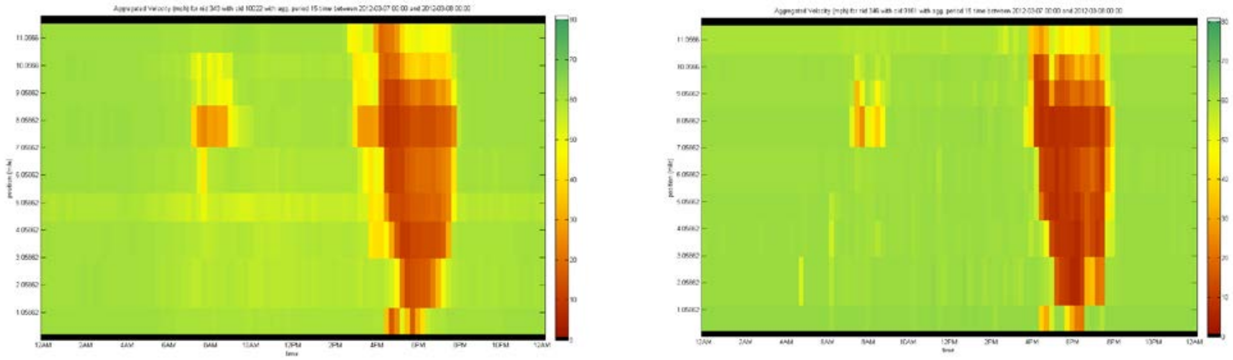


Figure 5-13: Average travel times presented as velocities on March 7 calculated from the reconstruction of vehicle trajectories using only loop data (left) and only probe data (right). The travel times are calculated over 15-minute intervals along each of nine routes between successive Bluetooth detectors.

Consistency among data sources. Figure 5-14 and Figure 5-15 reveal how the data sources compare, by showing the difference between the Bluetooth measured travel times and the estimated travel times from the trajectory reconstructions. Between the different data sources, it is clear that there is rough agreement over the times and locations of congestion.

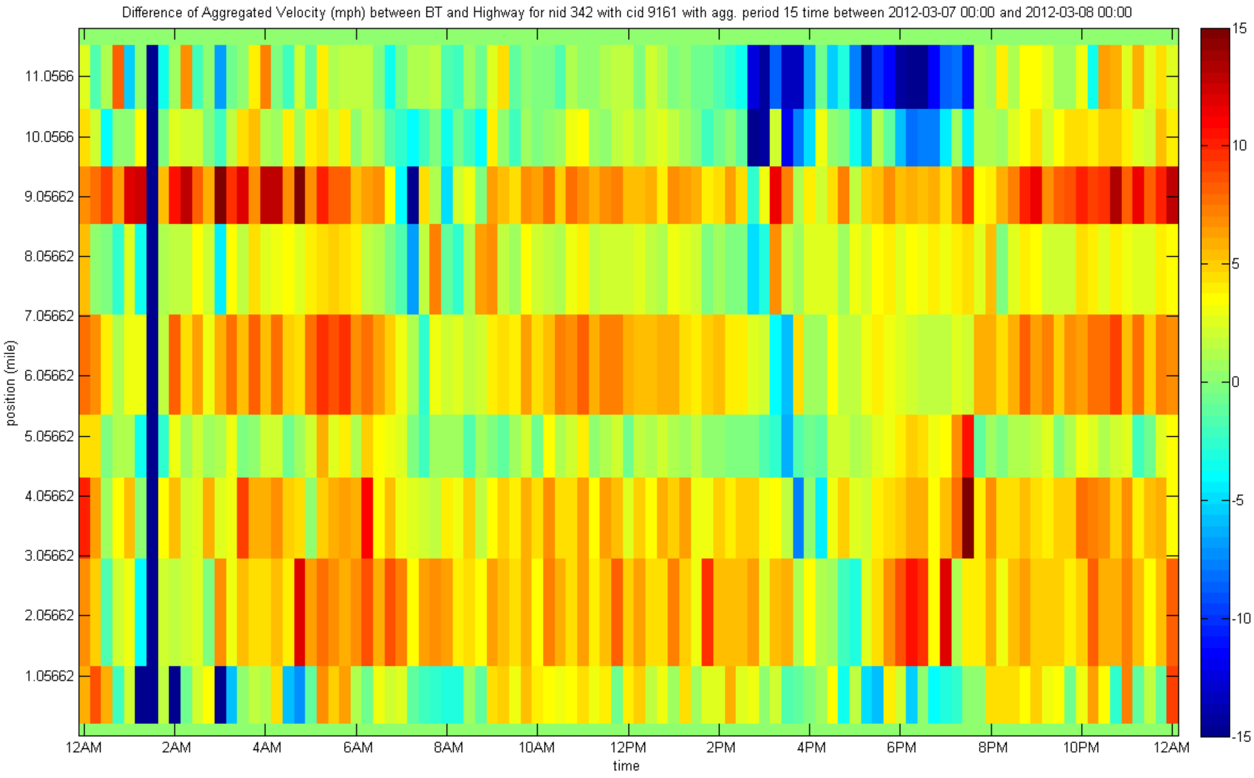


Figure 5-14: Difference calculated between Figure 5-11 and Figure 5-12, presented as velocities. The scale on the right ranges from -15 (15 mph over Bluetooth measurements) to +15 (15 mph under). Regions where the estimated travel times were overestimated are shown in red; underestimated travel times are shown in blue.

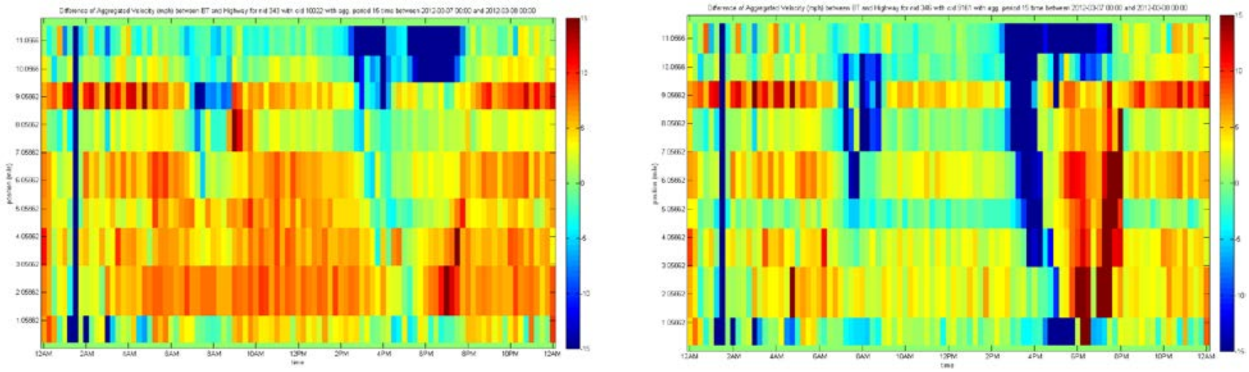


Figure 5-15: Difference calculated between Figure 5-13 loop data and Figure 5-11 (left), and between Figure 5-13 probe data and Figure 5-11 (right), presented as velocities. Regions where the estimated travel times were overestimated (underestimated) are colored red (blue). Notice that on this day, probe data does not capture well the onset of the afternoon congestion period.

5 DESCRIPTION OF METRICS

This section explains the metrics used to quantify the performance of the data fusion engine and the usefulness of probe data for the purpose of estimating travel times.

5.1 DEFINITIONS

MAPE

The mean absolute percentage error (MAPE) is used as a global metric calculated over selected regions:

$$MAPE = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\left|T_{hwy}(i,j)-T_{bt}(i,j)\right|}{T_{bt}(i,j)}$$

where *i* and *j* index time and space, respectively. N and M are, respectively, the number of five minutes per miles and the number of routes between Bluetooth sensors. For each of the experiments ten sensors were installed, yielding $M = 9$ routes in each direction. This metric quantifies the percentage error between the estimated travel time from the data fusion engine, T_{hwy}, and the actual measured travel times T_{bt}.

Often it is useful to plot the evolution of the MAPE over the course of one day. For this purpose, MAPE is also calculated at each fixed time interval indexed by *i*:

$$MAPE(i) = \frac{1}{M}\sum_{j=1}^{M}\frac{\left|T_{hwy}(i,j)-T_{bt}(i,j)\right|}{T_{bt}(i,j)}$$

PMATE

Another metric to help gain an intuitive feel for the results is the per mile time error (PMATE):

$$PMATE = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{L(j)}{\sum_{j=1}^{M}L(j)}\frac{\left|T_{hwy}(i,j)-T_{bt}(i,j)\right|}{L(j)}$$

$$= \frac{1}{N\sum_{j=1}^{M}L(j)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|T_{hwy}(i,j)-T_{bt}(i,j)\right|$$

where $L(j)$ is the length of the route indexed by *j*.

BTMAPE

A measure of the variability in Bluetooth measured travel times is also employed. This metric is the BTMAPE:

$$BTMAPE = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{S_{bt}(i,j)}{T_{bt}(i,j)}$$

where $S_{bt}(i,j)$ and $T_{bt}(i,j)$ are the standard deviation and mean, respectively, of the Bluetooth travel times observed in the time-space region indexed by *i* and *j*. The ratio of the standard deviation to the mean, $\frac{S_{bt}(i,j)}{T_{bt}(i,j)}$, is a normalized measure of dispersion known as the coefficient of variation.

BTPMATE

Like the BTMAPE, it is possible to compute the BTPMATE:

$$BTPMATE = \frac{1}{N\sum_{j=1}^{M}L(j)}\sum_{i=1}^{N}\sum_{j=1}^{M}S_{bt}(i,j)$$

CCE

The congestion classification error (CCE) is a metric used to quantify how well the data fusion engine estimates the presence of congestion. After defining a threshold P for the pace to be considered as congested or free-flow, CCE counts the percentage of space-time bins (see Figure 5-4) where the data fusion engine makes the same classification as the Bluetooth reference:

$$CCE = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|1_{\frac{T_{hwy}(i,j)}{L(j)}<P} - 1_{\frac{T_{bt}(i,j)}{L(j)}<P}\right|$$

The CCE is very sensitive to the threshold pace P. A pace of 1.5 min/mile (40 mph) was used to classify the bins.

CCEC

The congestion classification error in congestion (CCEC), while based on the CCE above, uses only those space-time bins that Bluetooth classifies as congested. From that set, we calculate the percentage of bins that are classified the same way (congested) by the data fusion engine.

5.2 METRIC OVER ONE DAY

This subsection describes the performance of the data fusion engine using several of the metrics defined above. March 7 is chosen as a representative day. Note, however, that on this day the Bluetooth data (Figure 5-11) shows an anomalous increase in travel times at 1:00 am that is not observed in the loop or probe data.

Special cases of performance metrics. When the data fusion engine is run with no data, it predicts free flow speeds everywhere. This configuration is used to understand what numerical values correspond to undesirable performance during periods of congestion. In addition, the inherent variability in travel times (quantified by BTMAPE or BTPMATE) corresponds to something of a “noise floor” for estimation performance.

Evolution of MAPE over one day. Figure 5-16 shows the evolution of MAPE(i) when the data fusion engine is run with PeMS data only (shown in cyan), probe data only (shown in pink), or all available data (shown in red). A MAPE of 0.4 or higher corresponds to predicting free flow when much of the freeway under study is in congestion. Although not shown, when run with no data the MAPE extends to 0.7 on this day. The variability in travel times is captured by BTMAPE and takes typical values around 0.2 or more during congestion and values around 0.12 during free flow.

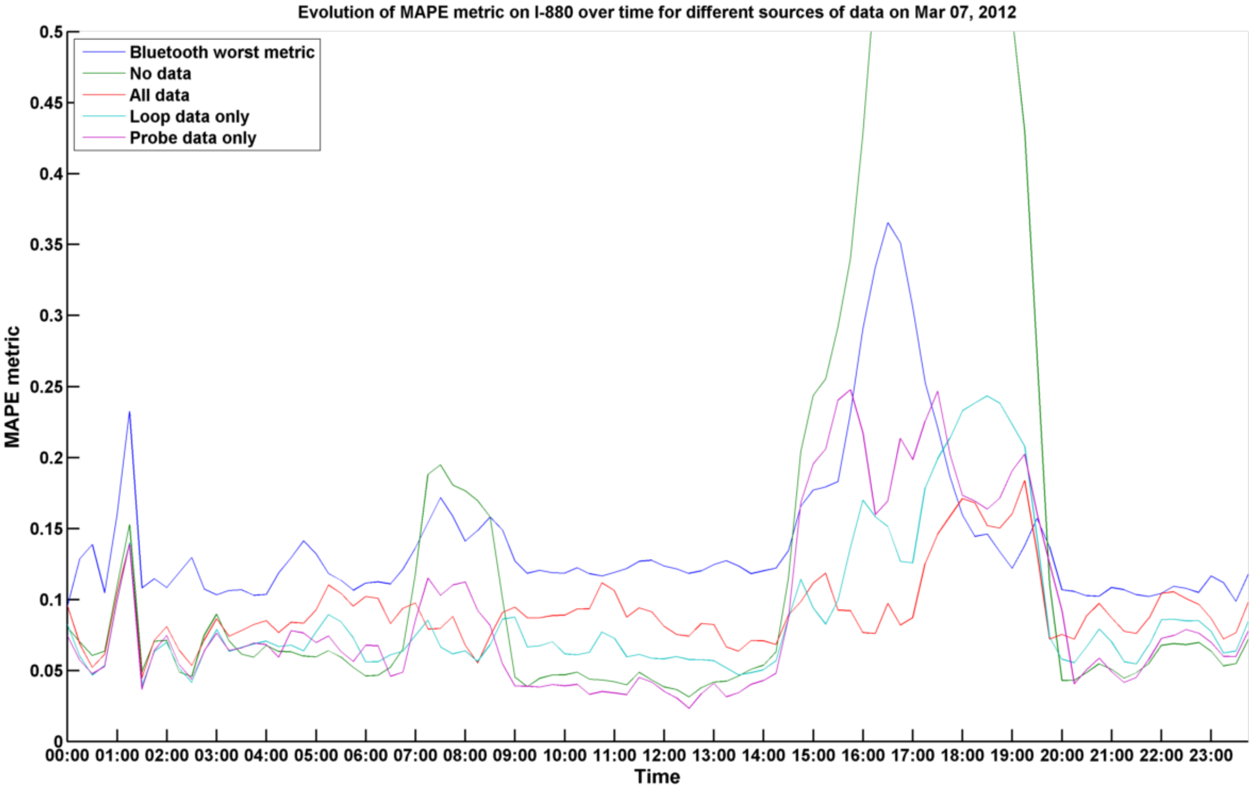


Figure 5-16: Evolution of MAPE and BTMAPE over the course of one day. The red, cyan, and pink lines correspond to MAPE using both PeMS and probe data, PeMS only, or probe data only, respectively. The green "No data" line corresponds to the MAPE when no data is available, and free flow is predicted everywhere. The "Bluetooth worst metric" in this plot is the BTMAPE and represents the variability in observed travel times.

Features of MAPE on this particular day:

- When only one data source (PeMS or probe) is used, the data fusion performance is worse than the BTMAPE during the last two hours of congestion.
- When all data is used, performance is almost always better than BTMAPE.
- Performance in congestion is better with PeMS data alone than it is with probe data alone.
- Performance in free-flow is slightly better with probe data alone than it is with PeMS data alone.
- The BTMAPE is unusually high near the beginning of congestion.

Evolution of PMATE over one day. The purpose of the PMATE metric is to get an intuitive sense of the per mile absolute time error. During the afternoon congestion period on this day, total travel time over the 12-mile freeway varies between about 36 and 43 minutes with an average of about 40 minutes. During congestion the per mile error peaks at about 40 seconds. Over 12 miles this would correspond to a worst case error of plus or minus 8 minutes, which is comparable to the observed variations in measured travel times (the BTPMATE). Not surprisingly, the features and time trends for the PMATE metric echo the findings of the MAPE metric, stated above. Although not shown, when run with no data the PMATE extends to 170 seconds per mile on this day.

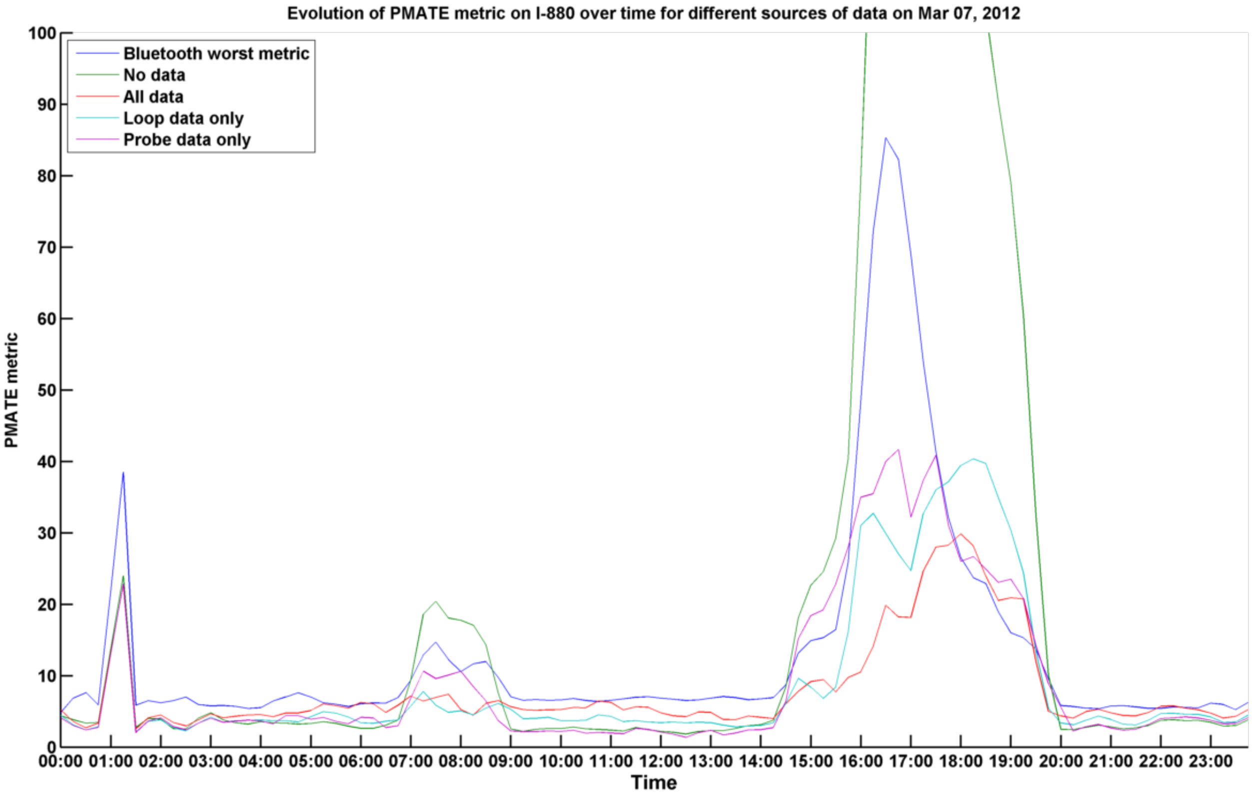


Figure 5-17: Evolution of PMATE and BTPMATE over the course of one day. The red, cyan, and pink lines correspond to PMATE using both PeMS and probe data, PeMS only, or probe data only. The green "No data" line corresponds to the PMATE when no data is available, and free flow is predicted everywhere. The "Bluetooth worst metric" in this plot is the BTPMATE and represents the variability in observed travel times.

Evolution of CCE over one day. The purpose of the CCE metric is to get an idea of how well the data fusion engine classifies regions of time-space as congested or not. Features of the congestion classification error (CCE) on this particular day are:

- The overall performance is best when all data are used. Furthermore, the classification is perfect during the morning peak. On average during the congested periods, the classification is 90% correct.
- Congestion covers the entire road around 5.30pm in the afternoon, and the classification is completely wrong when no data is available.
- Performance is better with PeMS data alone than it is with probe data alone.
- The CCE is unusually high near the beginning and end of congestion. This is consistent with the fact that the extremities of the queue are difficult to estimate in congestion.

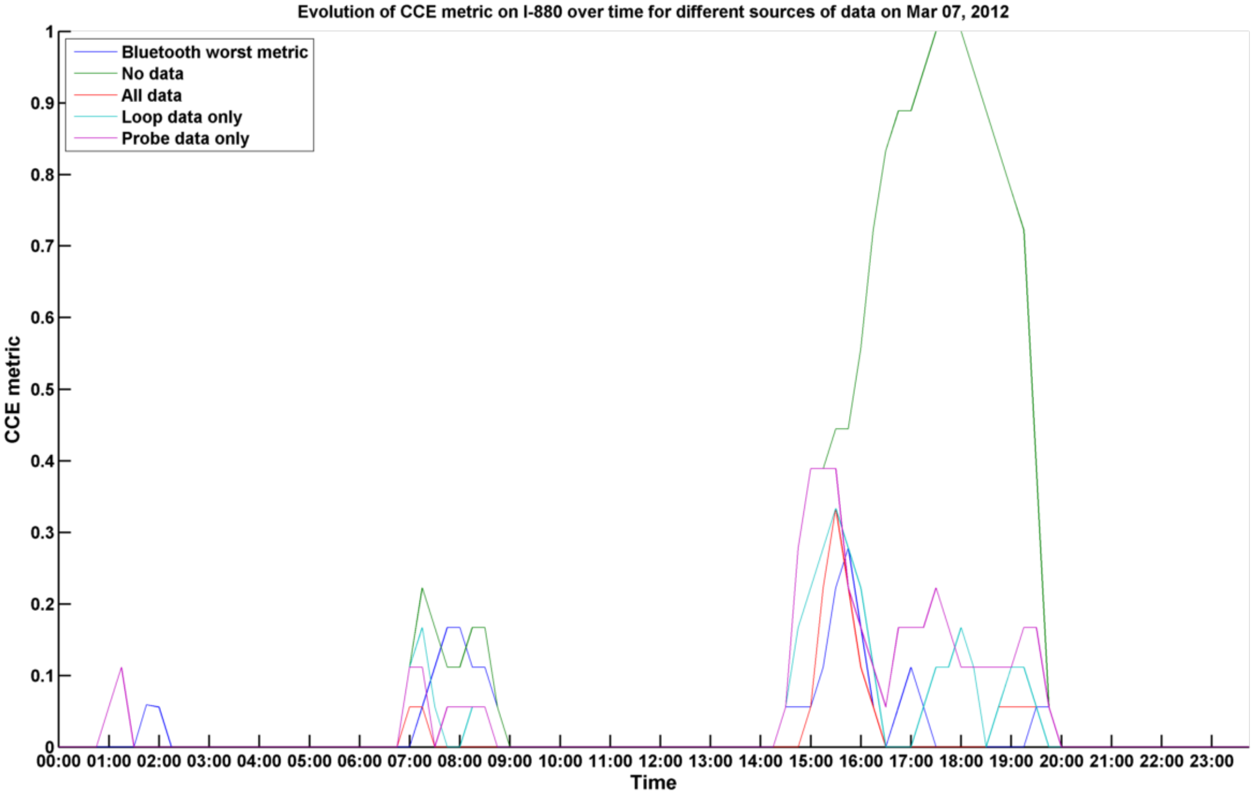


Figure 5-18: Evolution of CCE over the course of one day. The red, cyan, and pink lines correspond to CCE using both PeMS and probe data, PeMS only, or probe data only. The green “No data” line corresponds to the CCE when no data is available, and free flow is predicted everywhere.

6 DATA FUSION

This section reports the overall performance of the data assimilation algorithm and assesses the value of fusing data from multiple sources. We were able to adjust the amounts of loop and probe data and observe the results of using varying proportions of each.

There are several key points:

- Probe data is useful, even if it is sparse.
- When probe data and loop data are fused together, average travel times can be estimated at an accuracy within the bounds of driver variability.
- For the purpose of estimating travel times, probe data can be used as a substitute for data from loop detectors.

It is worth emphasizing that these results were obtained using probe data that was commercially available during the months of March and April of 2012. The reported quantities of probe data used here are representative only of each specific study site at that period in time.

6.1 LOOP STATIONS ON I-880

Figure 5-19 and Figure 5-20 show the loop stations that were active and provided valid data during the period of the experiment and on the section of highway under study.

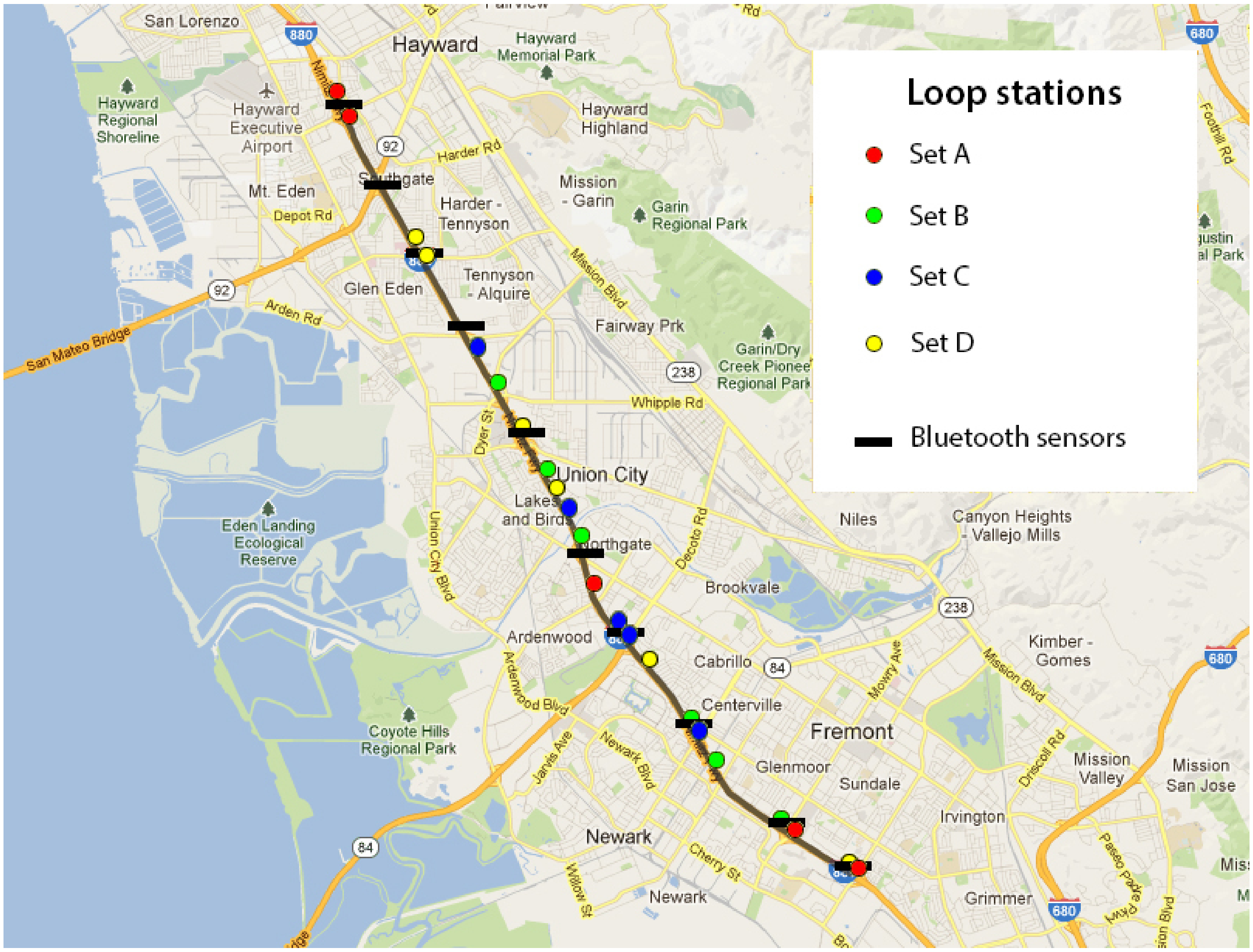


Figure 5-19: Loop stations used in the study of I-880

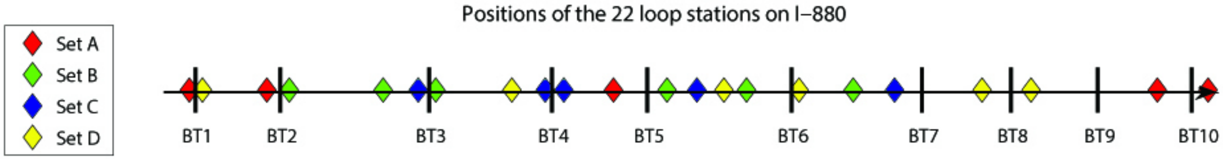


Figure 5-20: Loop stations used in the study of I-880, showing the same information as Figure 5-19, without the map

To include different percentages of the loop stations in the fusion calculations, we divided them into four sets, depending on their order of inclusion in the data fusion engine.

To include this:	We used this:
One quarter of the loop stations	Set A
Half the loop stations	Sets A + B
Three quarters of the loop stations	Sets A + B + C
All the loop stations	Sets A + B + C + D

The incremental addition of loop stations is shown in Figure 5-21:

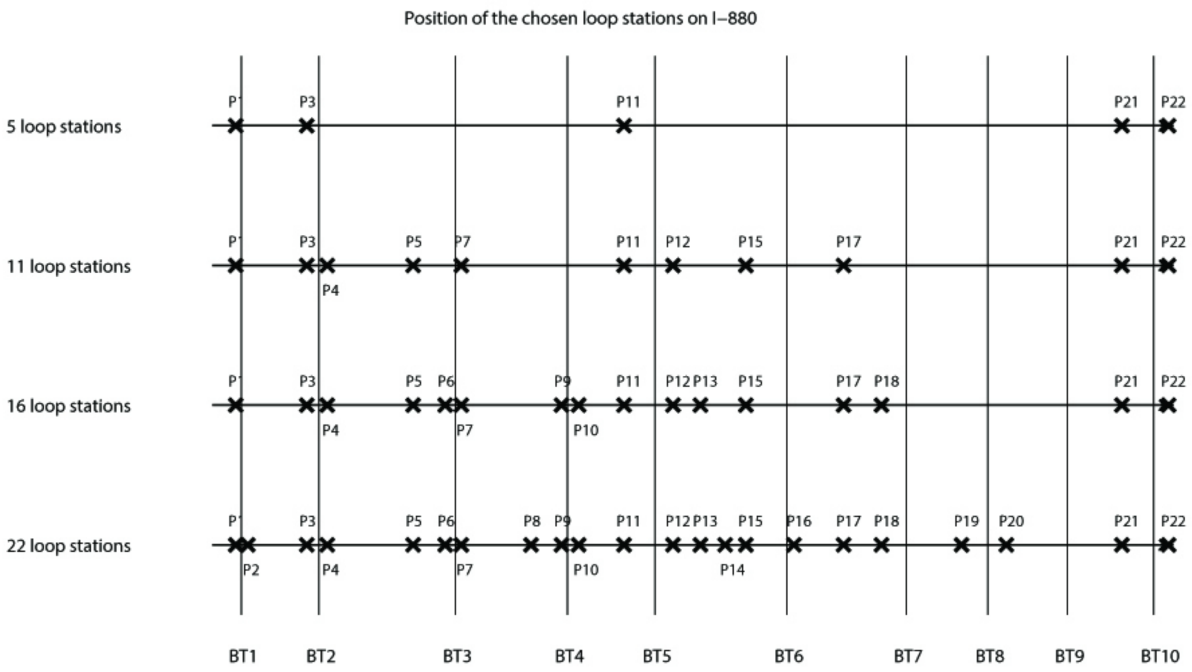


Figure 5-21: Different sets of loop stations on I-880 used in the data fusion engine. From top to bottom, loop stations included when using one quarter, half, three quarters, and all of the loop stations available and active during the period of the experiment.

A heuristic method was used to group loop stations into appropriate sets. The purpose was to obtain an intuitive and reasonably regular decrease in the error metrics when incrementing the number of loop stations. This “regular” decrease can be seen in Figure 5-22 when looking at the starting point of each line series. Therefore, a few model instances were run using data from a few combinations of loop stations. At each incremental step, the set giving the most regular decrease was chosen. A more precise and robust technique would be to randomly create multiple sub-samples of loop stations for each incremental fraction of loop stations, and compute the corresponding range of error metrics. This would be an area for further research.

6.2 METRIC VS. NUMBER OF UNIQUE PROBE DEVICES FOR FIXED FRACTIONS OF LOOPS

MAPE. Figure 5-22 shows that probe data can be used to improve or to supplement loop detector data for the purpose of estimating travel times. For each line series, the number of loop detector stations was fixed. The amount of probe data was adjusted using the method described in section 7.1.

Note that in the data subselection method described in section 7.1, unique probe devices are counted in 30-second time windows over the entire area of study. In this way, the notion of unique is intrinsically tied to the data assimilation time step, which is also 30 seconds. For greater readability, however, in the figures presented below and in later sections the horizontal axis has been normalized in both time and space such that the number of unique devices is presented as an average number of devices selected in ten 30-second data assimilation five minutes per miles, or five minutes, over an average mile of the area of study.

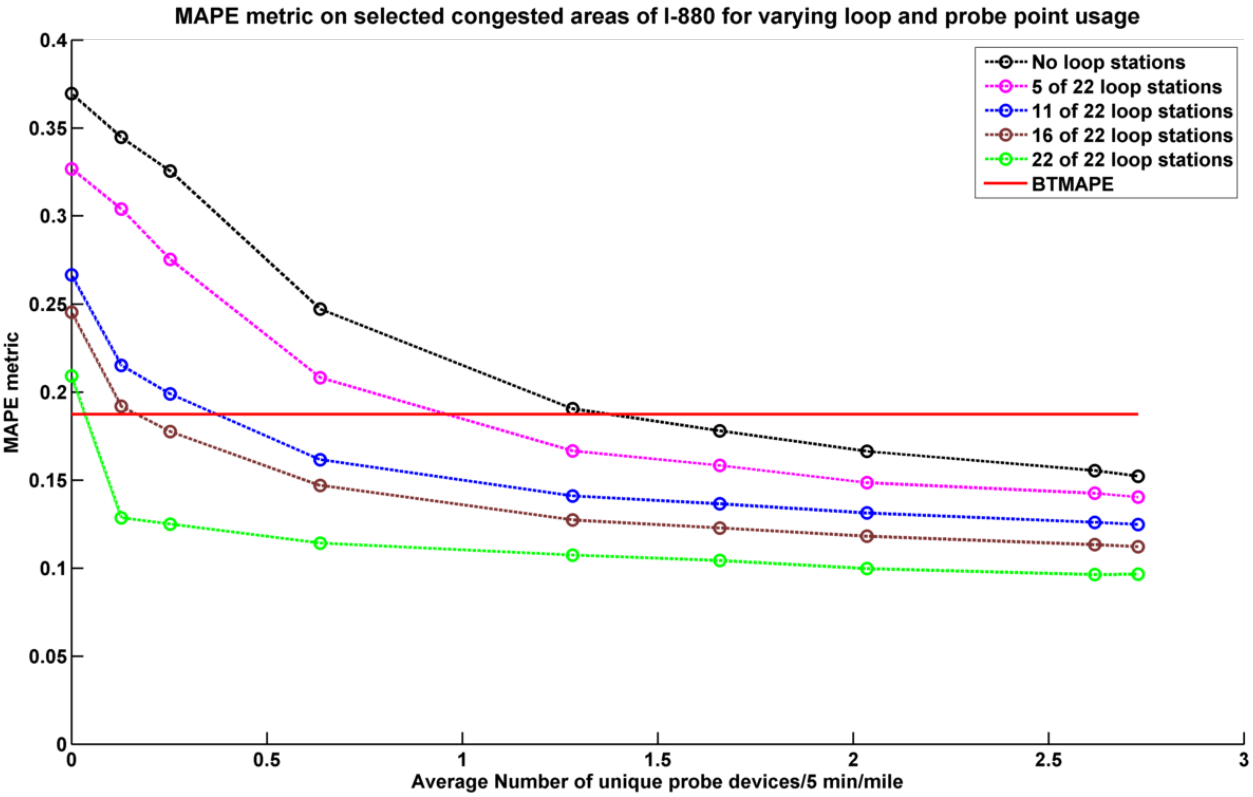


Figure 5-22: Global travel time MAPE plotted against the average number of unique probe devices used per five minutes per mile. Each dashed line corresponds to the performance achievable when fusing probe data with that from a different fraction of available loop detector stations. The red line corresponds to the "noise floor" of measured travel time variability.

For example, consider the case where all loop data is available, but no probes. The performance metric is 0.21 (21% difference from Bluetooth measurements). This same level of performance can be achieved by using 5 of the 22 loop detectors and having probe data corresponding to 0.7 unique probe devices per five minutes per mile. When using probe data alone (no loops), the same performance is reached with 1.2 unique probe devices per five minutes per mile.

Figure 5-22 also reveals the effect of using different sets of loop stations. When using data from 1.5 probe devices per five minutes per mile or less, the error is clearly reduced by adding 5, 11, and 22 loop stations, but the performance improvement between 11 and 16 loop stations is much smaller. This suggests that some loop stations add more value to the model than others and that their locations might be more significant than the minimal spacing between them. This would be an area for further investigation.

With the probe data in this example, improvement in the metric is most apparent when adding up to 1.5 unique devices per five minutes per mile. Adding more data beyond that shows a much more gradual improvement.

Probe flows, 85th percentile. For the business analysis of data fusion[4], we wanted to estimate the flow captured by probe data from the number of probe uniques per five minutes per mile. To provide a conservative value for probe flow to achieve the desired performance metric, we decided to compute and show the 85th percentile probe flow over the selected time periods (the areas of congestion identified in Figure 5-5 and Figure 5-6 where the global metrics were computed). By taking a high percentile, we essentially discard the time periods during which little or no data is collected, especially at night. Conversely, this metric is not affected by a few “lucky strikes” (selected hours during which a particularly high volume of vehicle data is collected, but that are not statistically representative).

To figure out the relationship between the two variables, we sampled the original data set several times by eliminating randomly chosen unique identifiers along with their data. This created 10 subsets with an increasing total number of probe uniques. For these samples, both the probe flow and the number of probe uniques per five minutes per mile were calculated. A linear model between the two variables provided a good fit: flow (veh/hour) = 9.32 * average number of uniques/5 min/mile (P-value < 1e-10, adjusted R-squared = 0.99), as illustrated in Figure 5-24. This technique enabled us to change the x-axis of the plot in Figure 5-22 from the average number of probe uniques/5 min/mile to the 85th percentile flow shown in Figure 5-23:

---

[4] For the business analysis, see the *Hybrid Data Roadmap* chapter of the Task Order 2 final report, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods*.

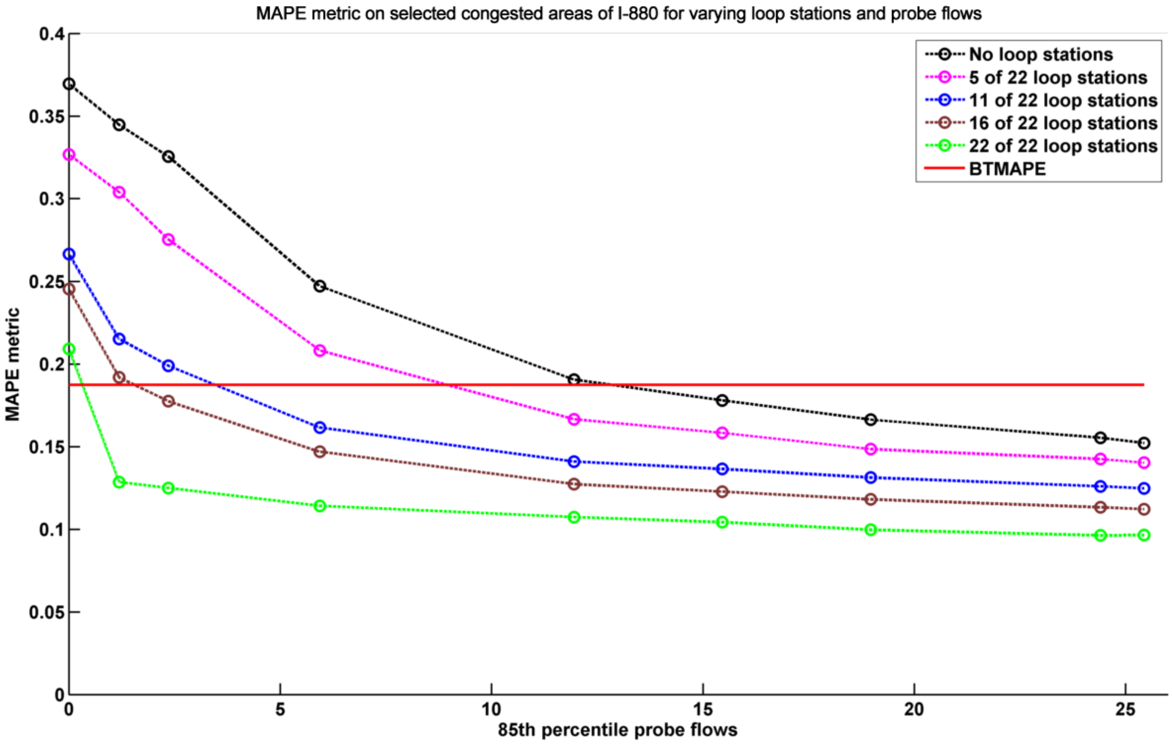


Figure 5-23: Contours of constant travel time estimation performance as a function of both number of loop detectors and 85th percentile flow (vehicles per hour). The lines represent the MAPE (percent discrepancy from Bluetooth measurements), with only probe data along the x-axis and only loop data along the y-axis.

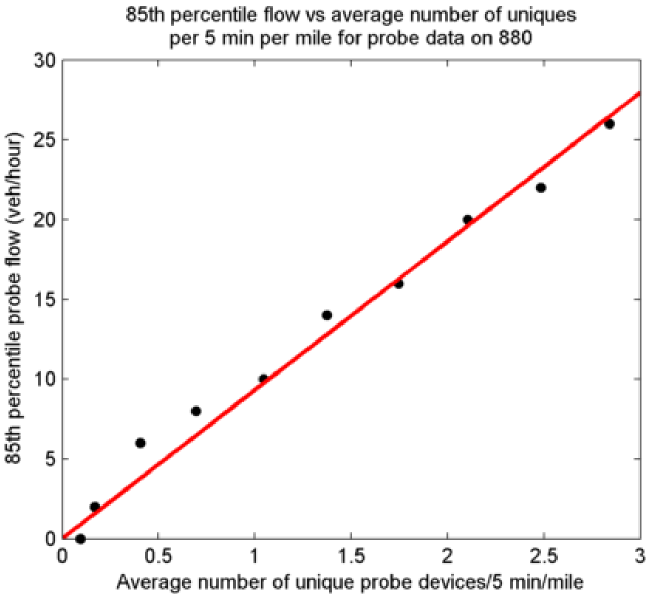


Figure 5-24: 85th percentile probe flow as a function of the average number of probe uniques/5 min/mile. The black points are the real values computed for 10 samples of the original data set. The red line is the result of a linear regression with a zero intercept. This very good fit justifies replacing the number of uniques per five minutes per mile with a linear approximation of the probe flow in other illustrations.

PMATE. Not surprisingly, Figure 5-25 yields the same conclusions as Figure 5-22, but this time using the PMATE metric. When running the model without data (resulting in free-flow speed everywhere), the metric is 57 seconds per mile. For the 12-mile section of freeway, it will incur in the worst case a 11.5-minute error in travel time estimation. When all loop data is available but no probes, the metric is 31 seconds per mile, incurring in the worst case a 6-minute error in travel time prediction. Adding probes makes it possible to lower the metric to 12 seconds/mile, which results in a 2.5-minute error in the worst case.

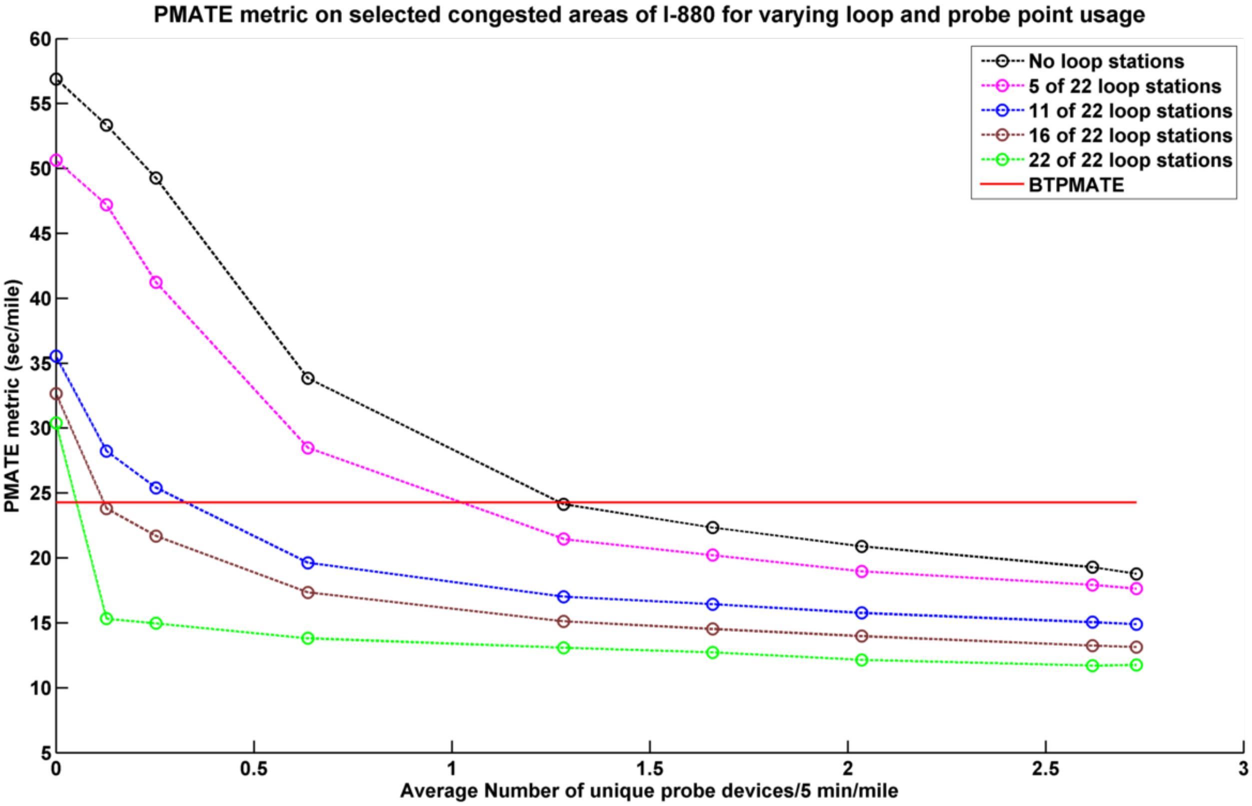


Figure 5-25: Global travel time PMATE plotted against the average number of unique probe devices used per five minutes per mile. Each dashed line corresponds to the performance achievable when fusing probe data with that from a different fraction of available loop detector stations. The red line corresponds to the “noise floor” of measured travel time variability.

CCEC. Estimates of congestion can also be improved with the addition of probe data. In Figure 5-26, for example, when using all the loop data but no probes, the data fusion engine shows a CCEC metric of 0.25 (a 25% difference from the Bluetooth measure of congestion). The same 0.25 metric is also achieved when using only half the loop stations and data from 1.3 unique probe devices per five minutes per mile.

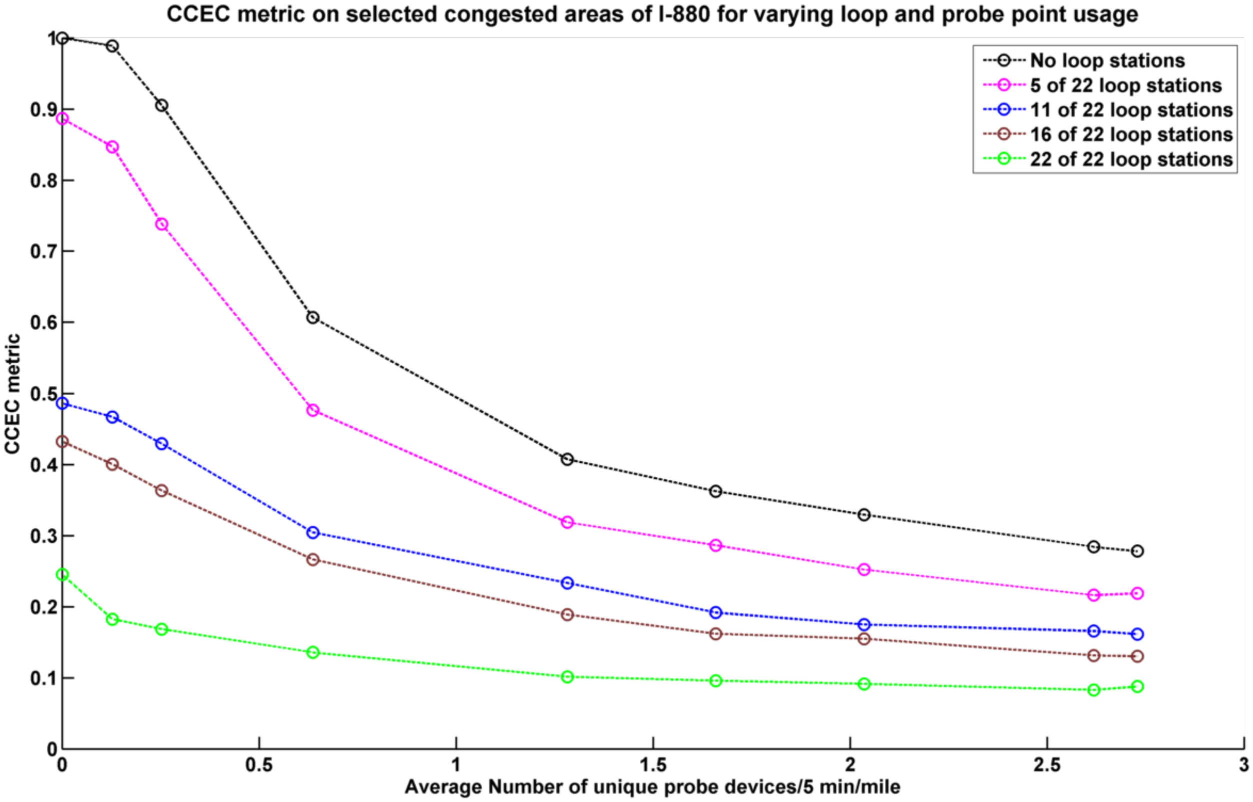


Figure 5-26: CCEC for congested areas of I-880, plotted against the average number of unique probe devices used per five minutes per mile. Each dashed line corresponds to the performance achievable when fusing probe data with that from a different fraction of available loop detector stations.

6.3 DATA FUSION ISOMETRICS

The data fusion isometrics presented here directly compare the predictive power of loop vs. probe data. For example, Figure 5-27 shows combinations of probe and loop data that could be used to achieve a given MAPE.

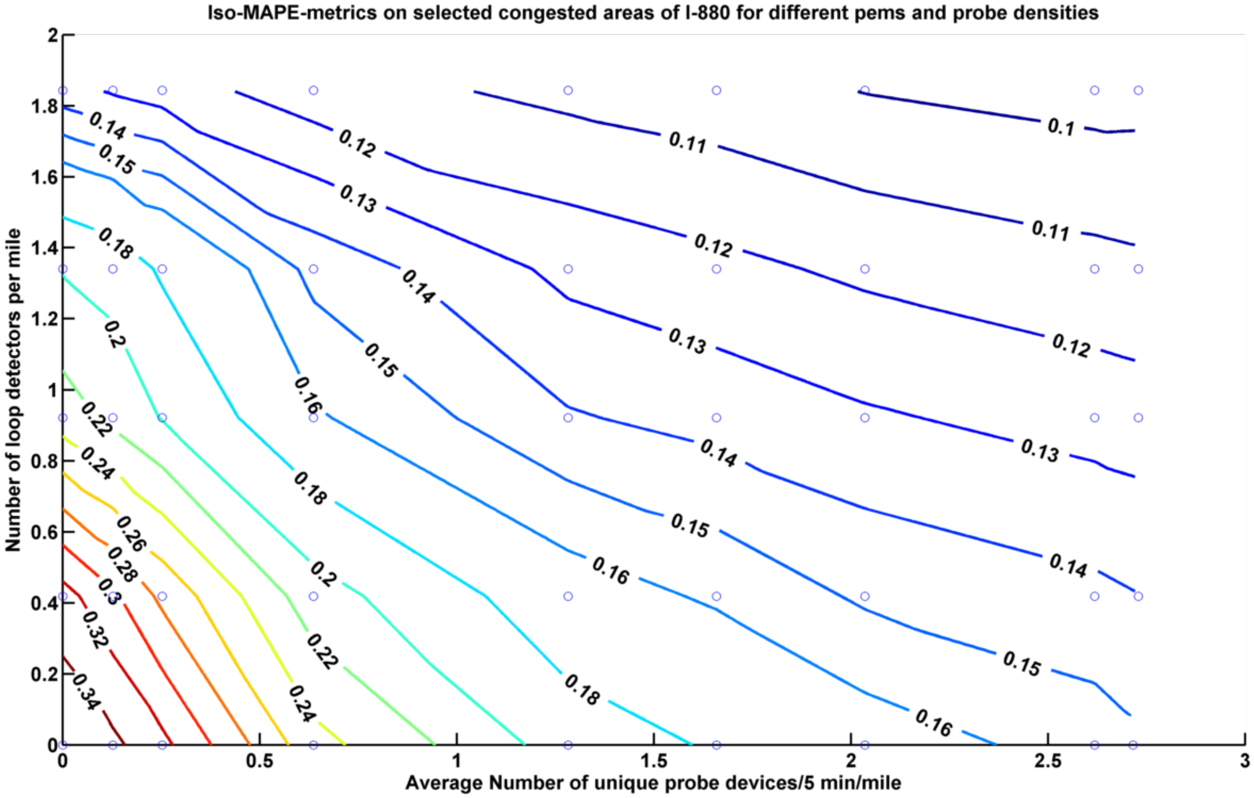


Figure 5-27: Contours of constant travel time estimation performance as a function of both number of loop detectors and average number of unique probe vehicles per five minutes per mile. The lines represent the MAPE (percent discrepancy from Bluetooth measurements), with only probe data along the x-axis and only loop data along the y-axis. The circles represent the actual points where the metric has been calculated. They form a grid and make it possible to generate the contours by interpolation.

On this plot, a MAPE of 0.18 (18%) is obtained with data from approximately 1.5 loop stations per mile with no probe data, or data from 1.6 probe devices per five minutes per mile with no loop data. Between those two endpoints, a range of loop/probe combinations along the contour could be used to achieve the 0.18 MAPE.

As with the MAPE metric described in section 6.2, we can also use the 85th percentile probe flows in the isometric comparison, shown in Figure 5-28:

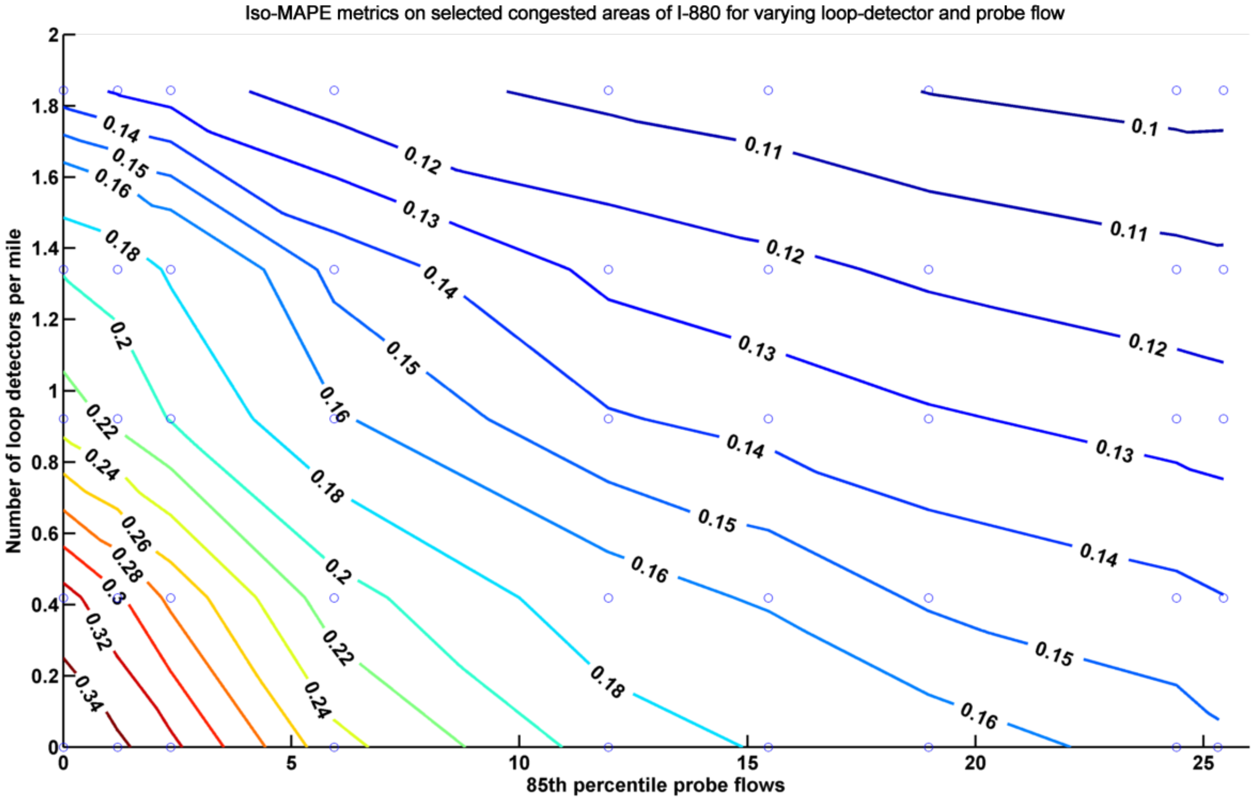


Figure 5-28: Contours of constant travel time estimation performance as a function of both number of loop detectors and 85th percentile flow (vehicles/hour). The lines represent the MAPE (percent discrepancy from Bluetooth measurements), with only probe data along the x-axis and only loop data along the y-axis. The circles represent the actual points where the metric has been calculated. They form a grid and make it possible to generate the contours by interpolation.

7 DATA QUALITY

This section explores secondary aspects of probe data that impact performance.

- **Sensitivity to quantity/quality.** The performance of the data fusion engine is found to be less sensitive to the raw quantity of data and more sensitive to the quality of that data. With probe data, it is better to have less frequent data from a larger number of unique vehicles (i.e., a high penetration rate) than it is to have more frequent data from a smaller number of unique vehicles (a high sampling rate).
- **Sampling rate.** We examined the temporal sampling rate of probe devices to see how frequently GPS data points were reported.
- **Penetration rate.** We examined the penetration rate of probe devices by comparing the flow of probe vehicles to all vehicles.

7.1 HIGH-FREQUENCY VS. LOW-FREQUENCY PROBE DATA

Two categories of probe data. Two categories of probe data were used for the analysis that follows:

- Type A probe data was taken from a relatively smaller population of vehicles. However, these vehicles supplied data points at a frequency greater than one point every 10 seconds (“high-frequency” data).
- Type B probe data was taken from a relatively larger population of vehicles. However, Type B vehicles provided data relatively rarely—often much less than once per minute (“low-frequency” data).

Two methods to subselect data. Two methods were used to subselect probe data:

- In the first method, a target was set for the number of unique devices during any data assimilation five minutes per mile (30 seconds) over the length of the whole area of study. Data from extraneous devices was discarded. Note that this target is sometimes not attained when not enough probe data points are available. This method was used to determine the sensitivity of the data fusion engine to a rough proxy for the probe vehicle penetration rate.
- In the second method, a target was set for the number of probe data points to be used during any data assimilation five minutes per mile over the length of the whole area of study without regard to the number of data points from the same probe vehicle. This second method was used to determine the sensitivity of the data fusion engine to the total quantity of data available.

For greater readability, in the figures presented below and in later sections the horizontal axis has been normalized in both time and space such that the number of unique devices is presented as an average number of devices selected in ten 30-second data assimilation five minutes per miles, or five minutes, over an average mile of the area of study.

Sensitivity to number of unique probes. Figure 5-29 shows the relationship between MAPE performance and the number of unique probes for type A and B data. The only real difference between the trends for A and B data is that type A data appears to have better performance, but this is expected because there is much more type A data. Indeed, as we can see in Figure 5-30 there are on average about ten times more type A data than type B data on the highway section of interest and during the selected time periods.

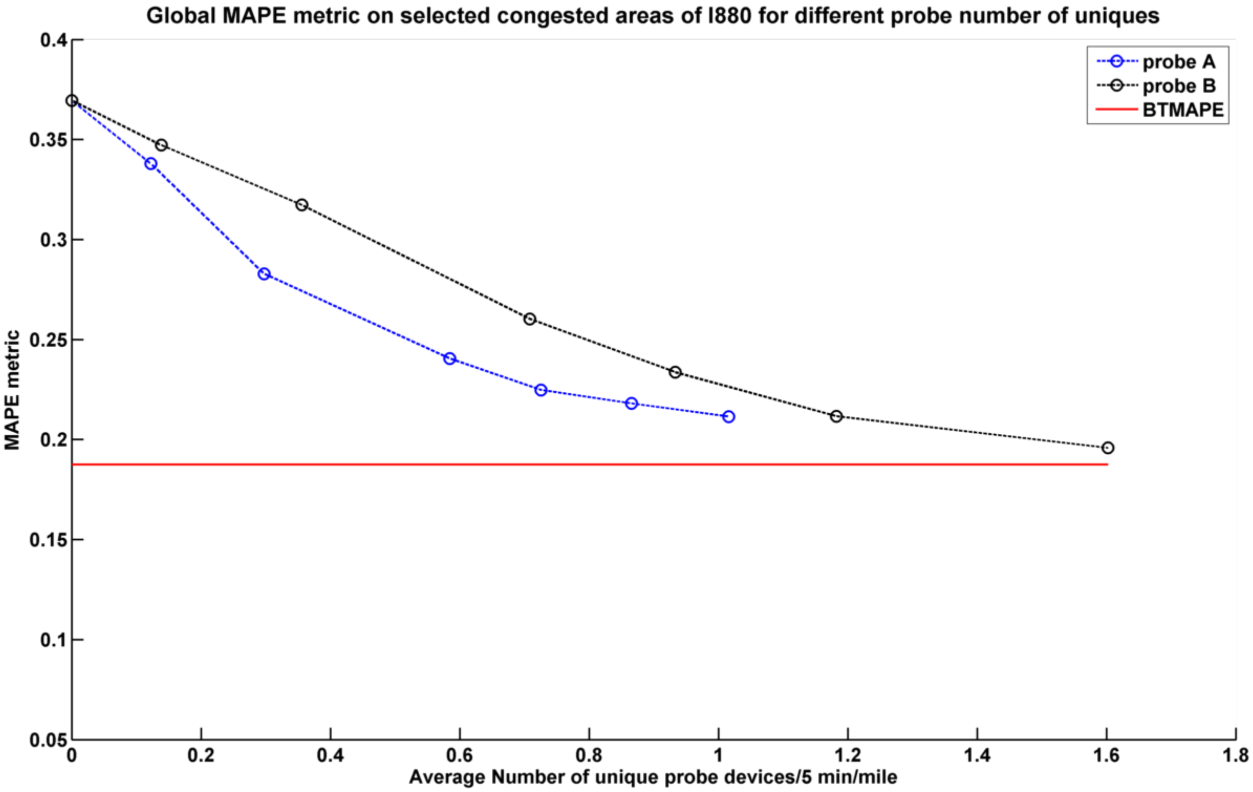


Figure 5-29: Relationship between MAPE performance and the number of unique probes for type A and B data (without PeMS data). The red line corresponds to the “noise floor” of measured travel time variability.

Sensitivity to number of data points. Figure 5-30 shows the relationship between MAPE performance and the number of probe data points for type A and B data. It can be seen that it takes more than ten times as much type A data as type B to achieve the same performance.

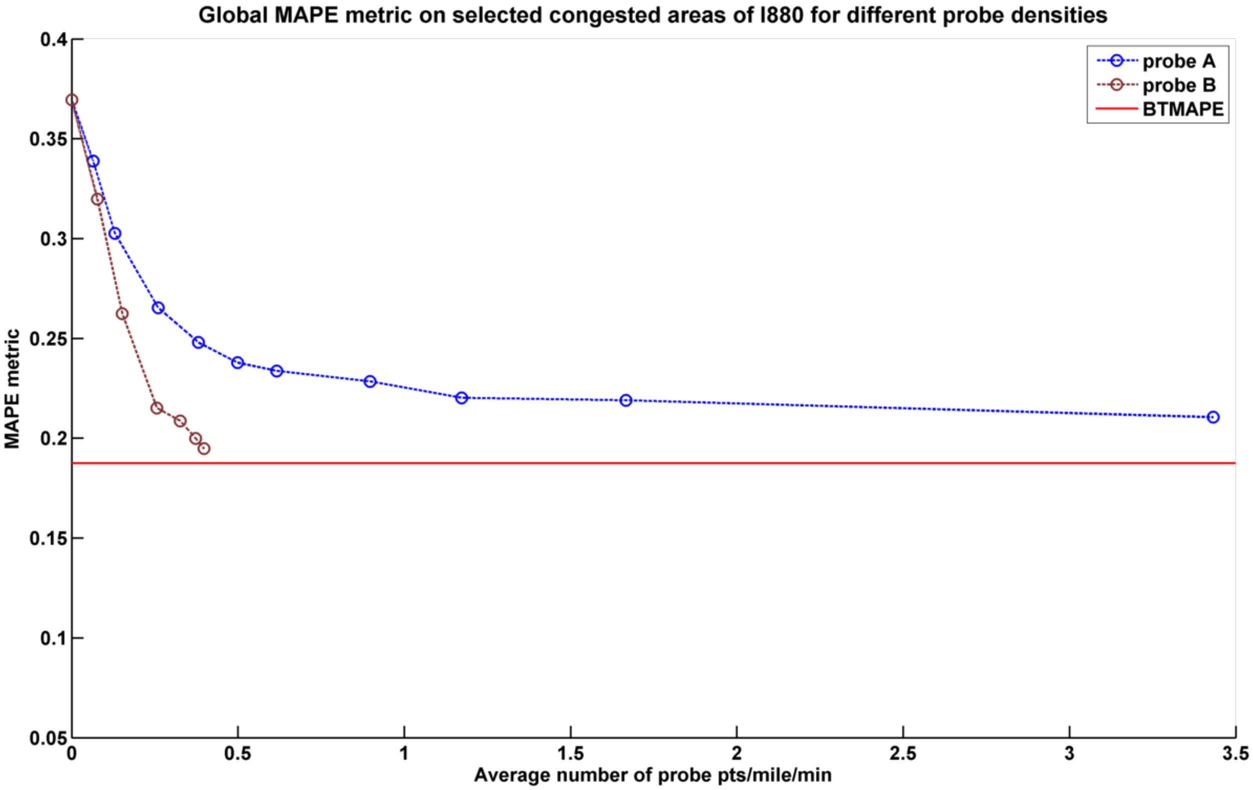


Figure 5-30: Relationship between MAPE performance and the number of probe data points for type A and B data (without PeMS data). The red line corresponds to the "noise floor" of measured travel time variability.

Total number of unique probes. Figure 5-31 shows the relationship between MAPE performance and the total number of unique probes for type A and B data combined. Using only probe data, the MAPE metric reaches the “noise floor” of measured travel time variability with fewer than 1.5 unique probe devices per five minutes per mile.

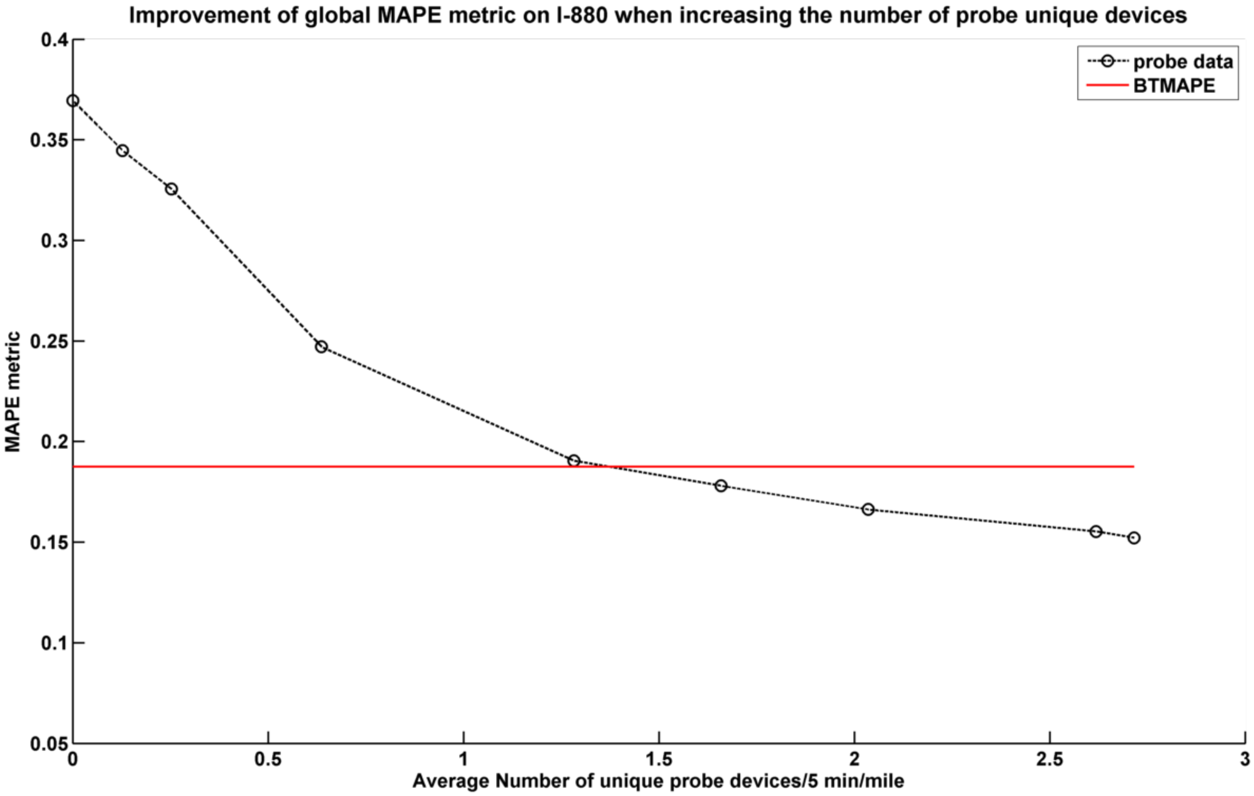


Figure 5-31: Relationship between MAPE performance and the total number of unique probes for both type A and B data (without PeMS data). The red line corresponds to the “noise floor” of measured travel time variability.

Total number of probe data points. Figure 5-32 shows the relationship between MAPE performance and the total number of probe data points for type A and B data combined. Using only probe data, the MAPE metric reaches the "noise floor" of measured travel time variability with an average of 0.5 probe points per mile per minute.

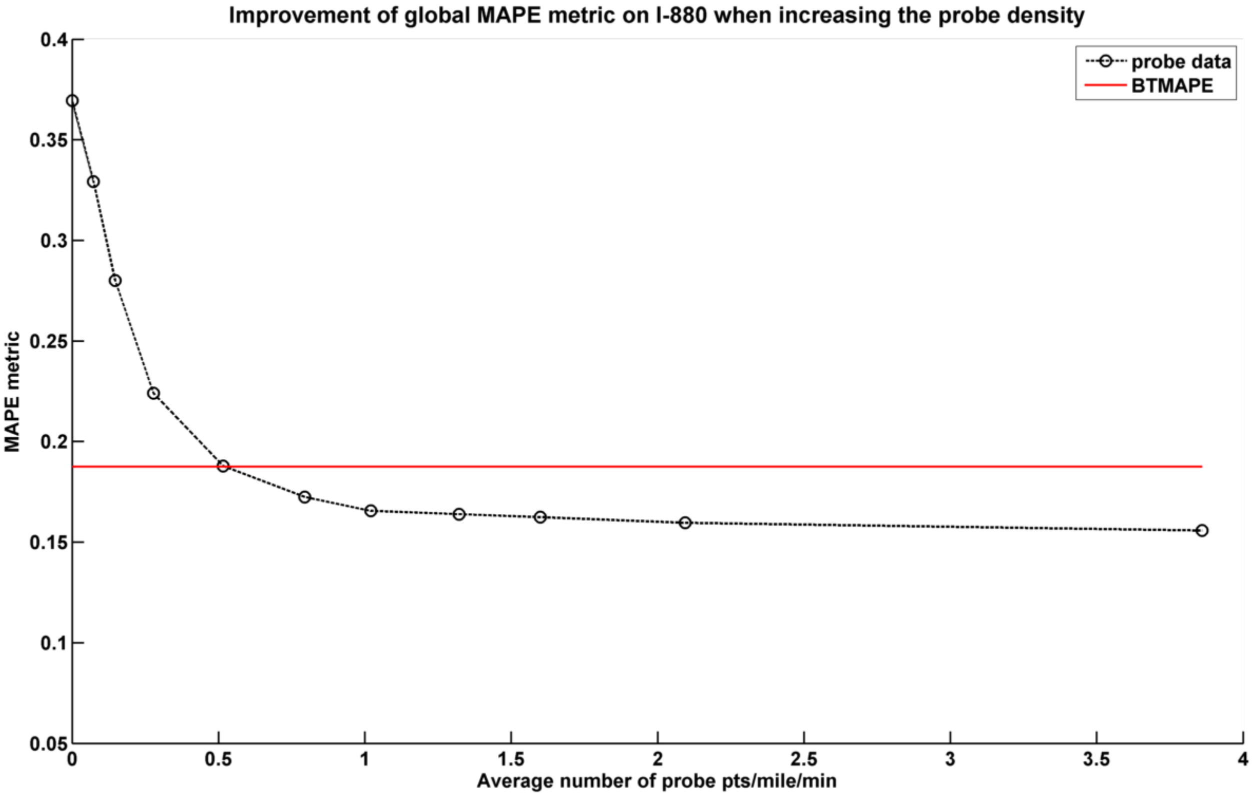


Figure 5-32: Relationship between MAPE performance and the total number of probe data points for both type A and B data (without PeMS data). The red line corresponds to the "noise floor" of measured travel time variability.

7.2 SAMPLING RATE

We evaluated the sampling rates of probe devices for the two categories of probe data along the 12-mile section of I-880 northbound over two weeks. The contrast between Probe A (high frequency) and Probe B (low frequency) can be clearly seen in the cumulative distribution function (cdf) shown in Figure 5-33:

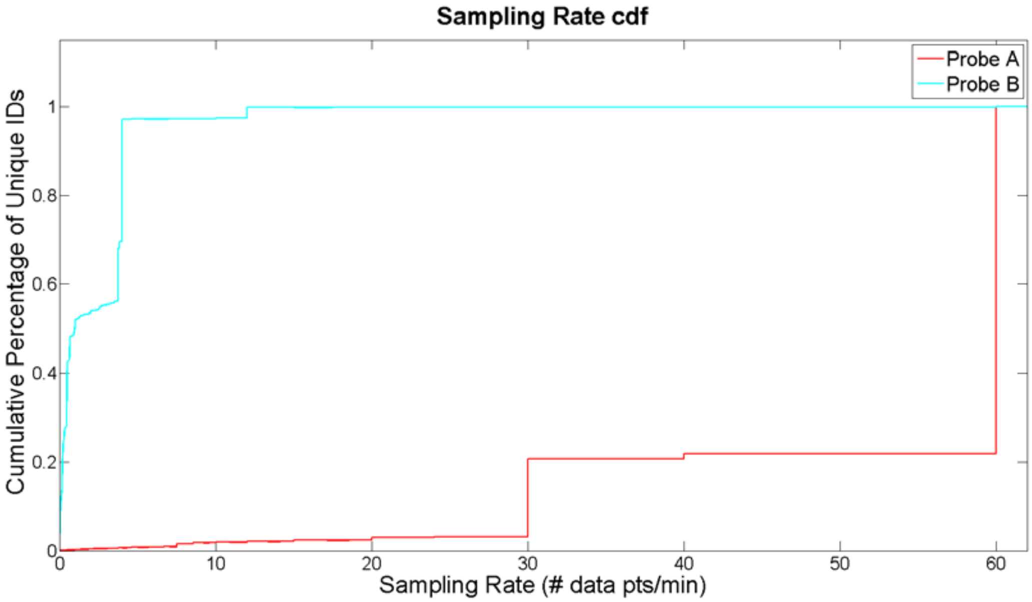


Figure 5-33: Comparison of Probe A and Probe B sampling rate cumulative distributions

The Probe A distribution shows a minimal percentage of unique devices with a sampling rate below 30 data points/minute, then more at 30, and a dramatic jump at 60 data points/minute. The Probe B distribution quickly rises at 2.5 data points/minute and then again at 5 data points/minute, with little change after that.

7.2.1 METHODOLOGY

To compare sampling rates, we use the Median Time Difference, as follows:

- Select all data points with probability[5] over 0.7 and group them by unique IDs.
- For each ID with more than one data point, take the time difference between subsequent data points.
- For each ID, take the inverse of the median time difference as its sampling rate.
- Plot the number of unique IDs versus sampling rates (number of data points per minute).

Note that for Probe B data, all time differences longer than two hours are discarded before finding the median time difference. This is because some IDs appear regularly (every day), which may bias the results.

7.2.2 RESULTS FOR PROBE A DATA

Figure 5-34 shows the results for Probe A data, using a log base 10 scale:

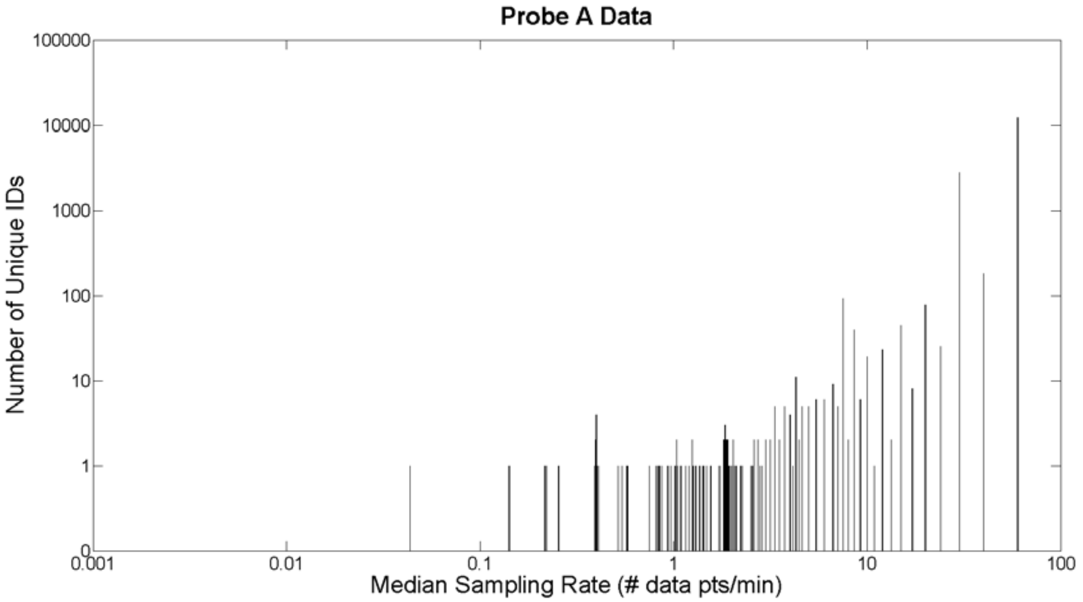


Figure 5-34: Median sampling rates for Probe A data

---

[5] This is the probability the data point is actually on the highway under study, rather than at a nearby location, as determined by the path inference filter (PIF). The path inference filter is described in the final report for Task Order 2, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods*.

- Total number of qualifying unique IDs: 15,751 (98.3% of total selected unique IDs)
- 1.7% of total selected unique IDs appear only once.
- Peaks: 7.5pts/min, 20pts/min, 30pts/min, 60pts/min—i.e., IDs that give a data point every 8 seconds, 3 seconds, 2 seconds, and 1 second (1 Hz)

Table 8 specifies the sampling rate intervals and corresponding percentages.

Table 8: Sampling rate intervals and corresponding percentages of probe devices for Probe A data

Sampling Rate (# data pts/min)	Percentage (%)	Cumulative Percentage (%)
(0–5)	0.75	0.75
[5–12)	1.22	1.97
[12–25)	1.14	3.11
[25–35)	17.66	20.77
[35–45)	1.15	21.92
[45–60]	78.08	100

7.2.3 RESULTS FOR PROBE B DATA

Figure 5-35 shows the results for Probe B data, using a log base 10 scale:

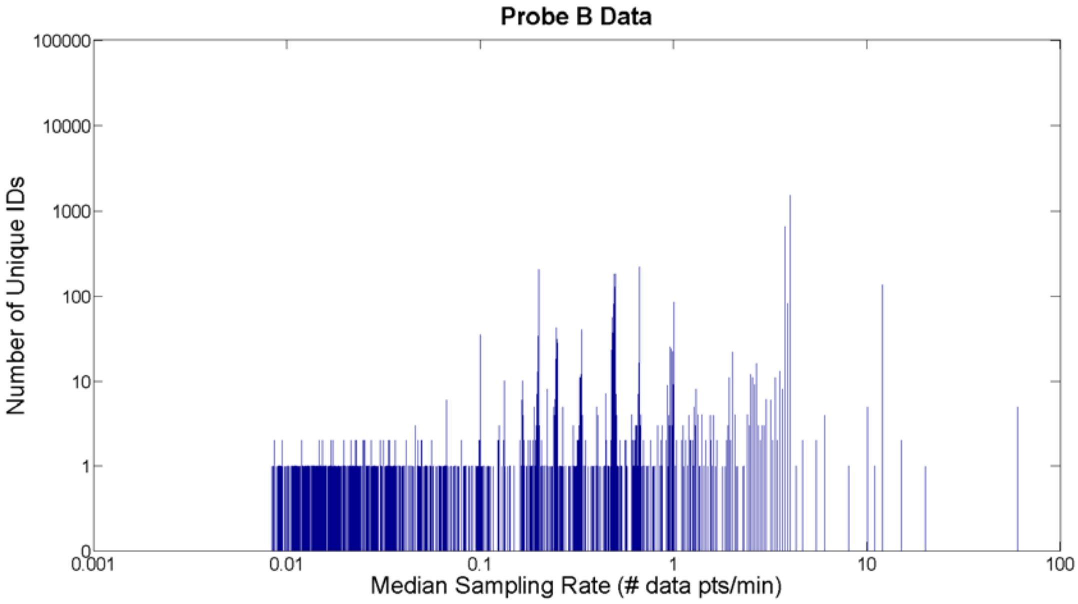


Figure 5-35: Median sampling rates for Probe B data

- Total number of qualifying unique IDs: 5,424 (78.36% of total selected unique IDs)
- 1,033 (14.92%) of total selected unique IDs appear only once
- 465 (6.72%) of total selected unique IDs do not have between data point time difference shorter than 2 hours
- Peaks: 0.5pt/min, 4pts/min, 12pts/min i.e. IDs that give a data point every 2 minutes, every 15 seconds and every 5 seconds

Table 9 specifies the sampling rate intervals and corresponding percentages.

Table 9: Sampling rate intervals and corresponding percentages of probe devices for Probe B data

Sampling Rate (# data pts/min)	Percentage (%)	Cumulative Percentage (%)
(0–2.5)	54.28	54.28
[2.5–5.5)	42.92	97.2
[5.5–13)	2.65	99.85
[13–35)	0.06	99.91
[35–55)	0	99.91
[55–60]	0.09	100

7.2.4 SUMMARY RESULTS FOR ALL SITES

Figure 5-36 and Figure 5-37 show the percentage of unique IDs for each sampling rate interval for I-880, I-15 Ontario, and I-15 Victorville for Probe A and Probe B data. The results clearly echo the findings for I-880 in Figure 5-33, Figure 5-34, and Figure 5-35, showing a high-frequency sampling rate for Probe A and a low-frequency rate for Probe B. Moreover, based on an initial assessment, it appears that the sampling rates of feeds in the different geographical zones are similar. The figures use two weeks of data from I-880 and I-15 Ontario and two months of data from I-15 Victorville.

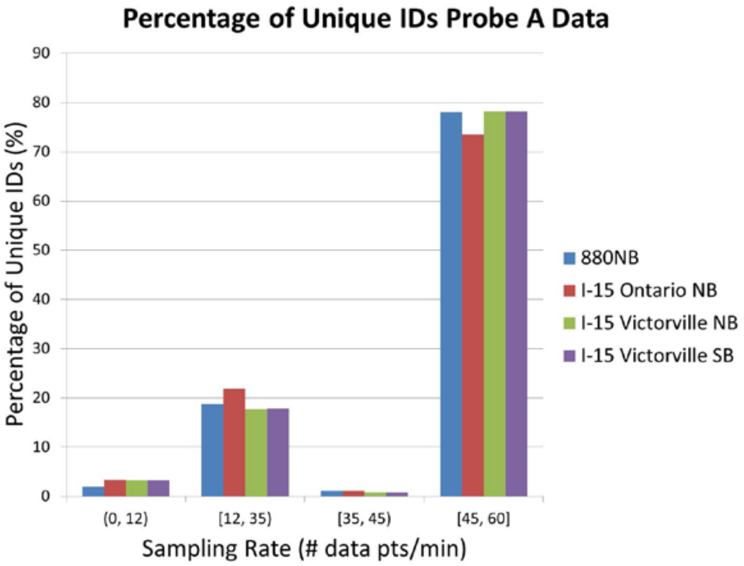


Figure 5-36: Summary results for Probe A data

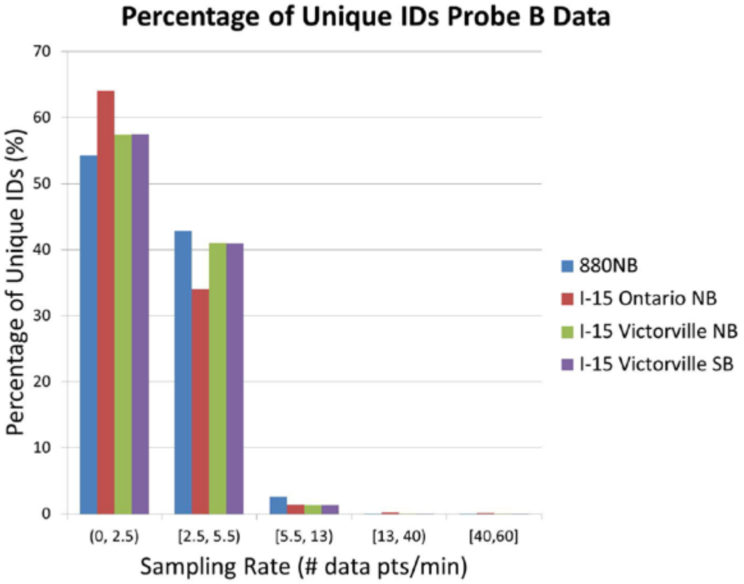


Figure 5-37: Summary results for Probe B data

7.3 PENETRATION RATE

Penetration rate provided another basis for comparing the two categories of probe data (Probe A, frequent data points from a relatively smaller population of vehicles; and Probe B, infrequent data points from a relatively larger population of vehicles).

Methodology. In order to find the penetration rate of probe devices, we need to compare the flows of probe vehicles to that of all vehicles. To do this, PeMS stations were chosen along the freeway, and we counted the number of probe devices that reported points before and after crossing the PeMS station. Since the PeMS station records flows of cars that cross it, this gives us flows across the same point in the highway. By dividing the flow of probe devices by the flow recorded by PeMS, a number is obtained that indicates the penetration rate of probe data.

Using the same highways and date ranges as the Bluetooth experiments, we retrieved data from three PeMS stations near the middle of the highway section, counting the probe flows and PeMS flows across each. We aggregated results for I-880 northbound over March 3 to March 17, and I-15 Ontario northbound over April 1 to April 12. (No loop detectors were active on I-15 Victorville during the experiment.)

Results. We found that the flows of both probe vehicles and all vehicles varied greatly between weekdays and weekends. On weekdays, rush hour in the morning and afternoon created a two-peaked shape on our plots, while on weekends there was only one peak in the afternoon, and it was not as high. Because of this, we grouped the data into weekdays and weekends and analyzed it separately for each of the groups.

7.3.1 I-880 WEEKDAYS

Figure 5-38 and Figure 5-39 show averages of probe and PeMS flow and the corresponding penetration rates for the I-880 site on weekdays. Probe B seems to yield a much higher flow than probe A, which is consistent across most of these findings. Probe B data also seems to peak slightly earlier than most traffic recorded by PeMS and yields the highest penetration rates at those times, around 0.8–1.0%. At other times, the Probe B penetration rate falls within the 0.2–0.7% range. Probe A flow seems to be relatively consistent throughout the day, but due to its significantly lower volume it never surpasses much more than a 0.1% penetration rate.

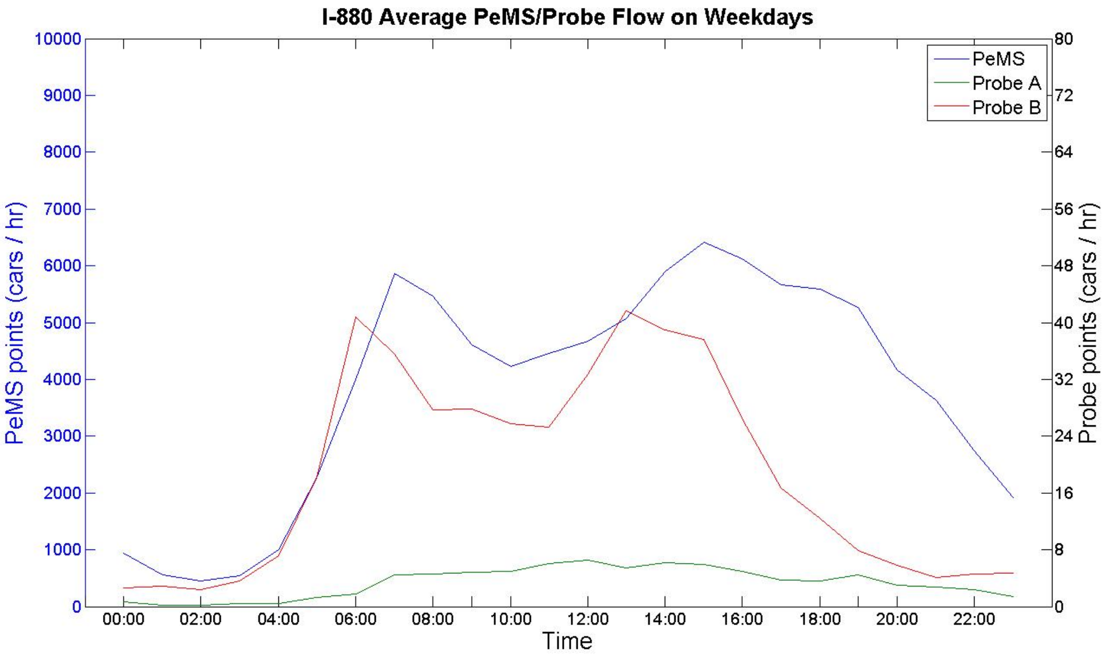


Figure 5-38: Average PeMS/probe flow on I-880 weekdays

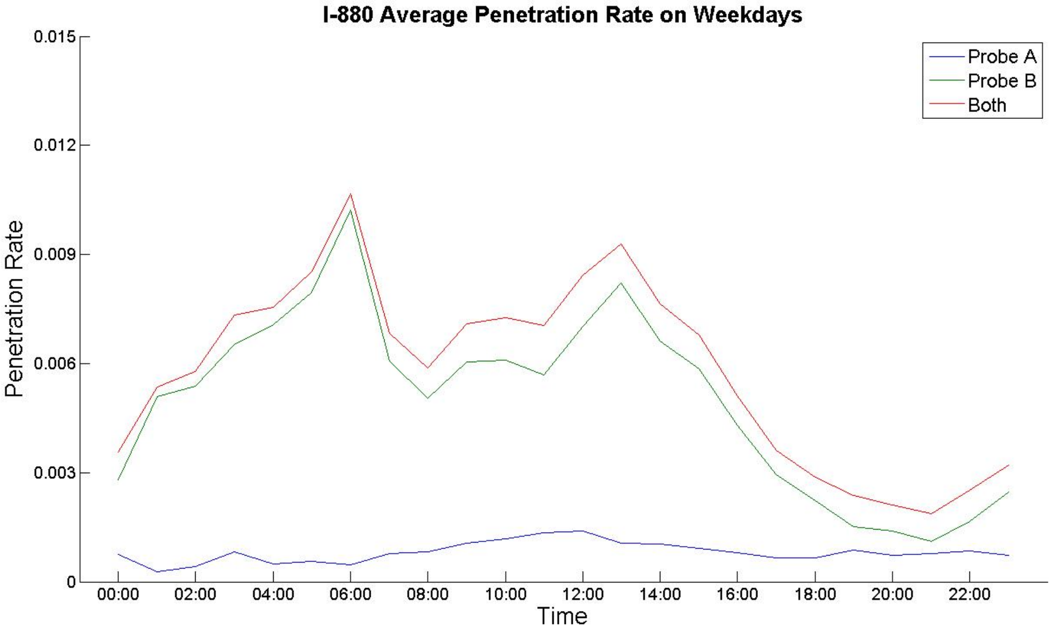


Figure 5-39: Average penetration rate on I-880 weekdays

Figure 5-40 and Figure 5-41 also show probe flow, PeMS flow, and penetration rate for the I-880 site, but instead of averaging the data across all PeMS stations and days, we now take the 85th percentile of the data. We use this to represent a "good day" for each provider (when a substantial amount of data is being reported). The trends are mostly the same, although we see that the probe flow benefits more from taking the 85th percentile than does the PeMS flow, and so we see consistent penetration rates in the range of 0.8–1.2%, which are higher than before.

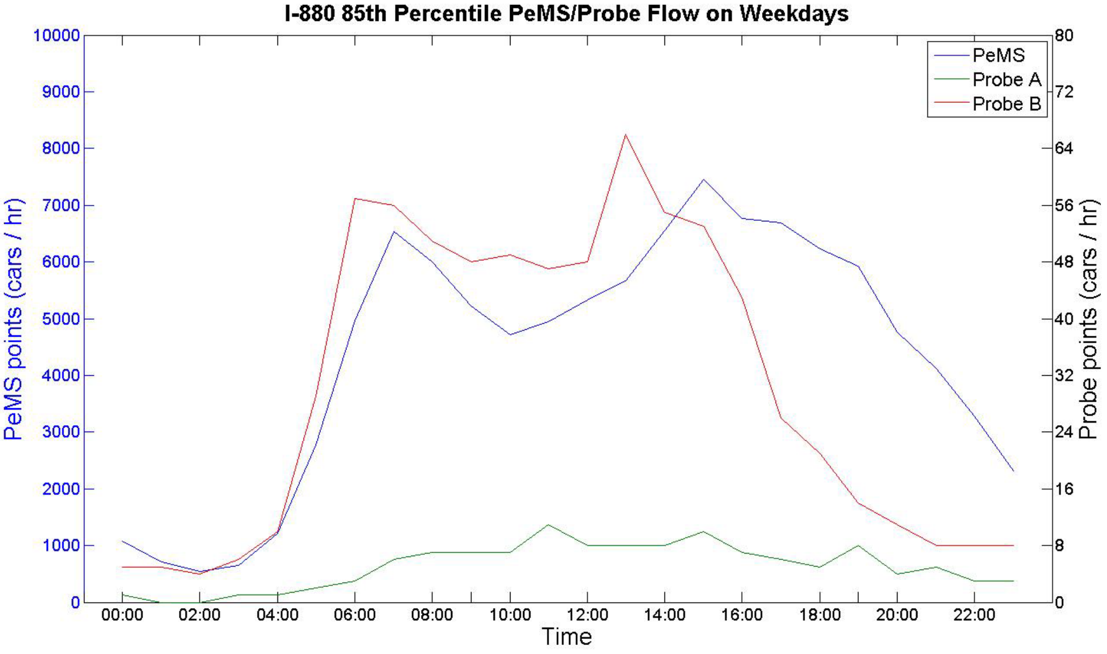


Figure 5-40: 85th percentile PeMS/probe flow on I-880 weekdays

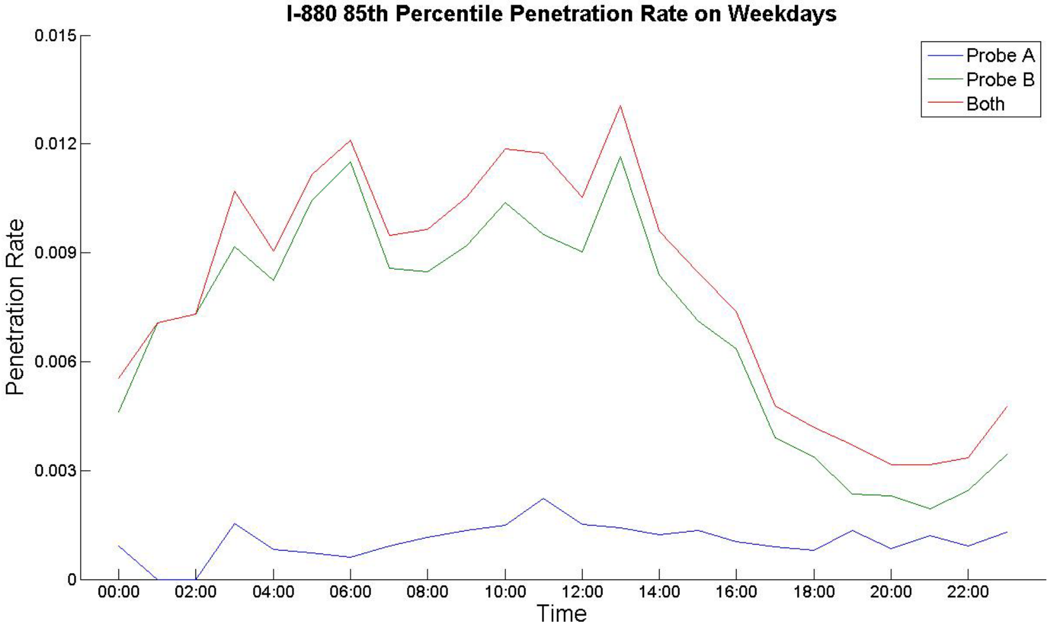


Figure 5-41: 85th percentile penetration rate on I-880 weekdays

7.3.2 I-880 WEEKENDS

Figure 5-42 shows average flows, again for the I-880 site, over weekend days. It seems that flows are lower overall, with the drop-off especially significant for Probe B. This reduces the penetration rates in Figure 5-43 to the 0.1–0.6% range.

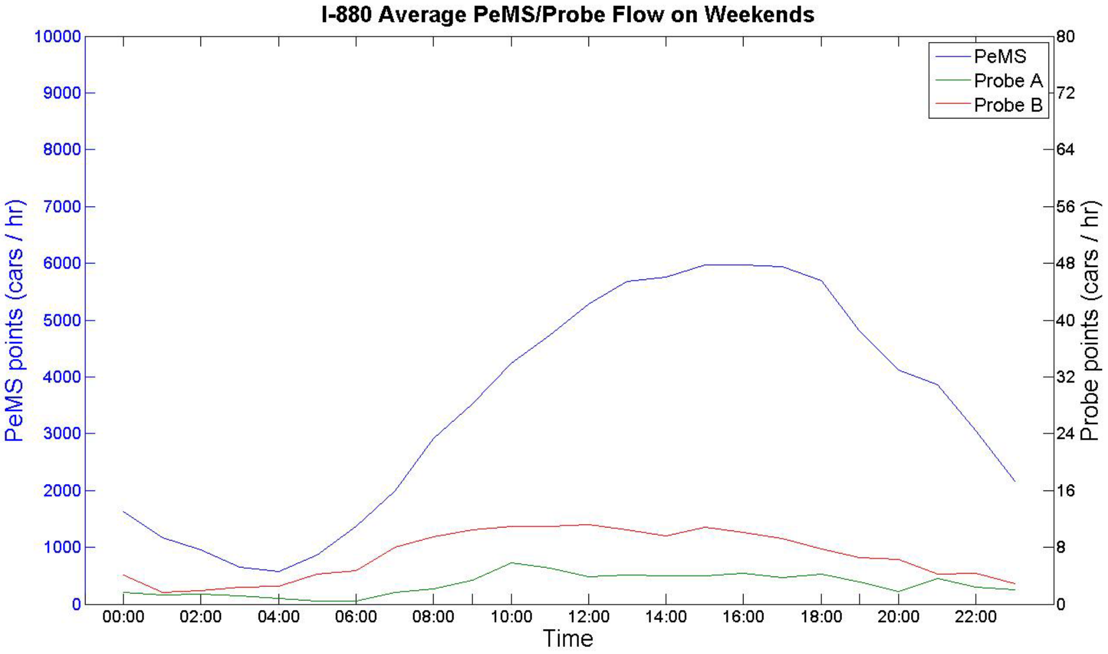


Figure 5-42: Average PeMS/probe flow on I-880 weekends

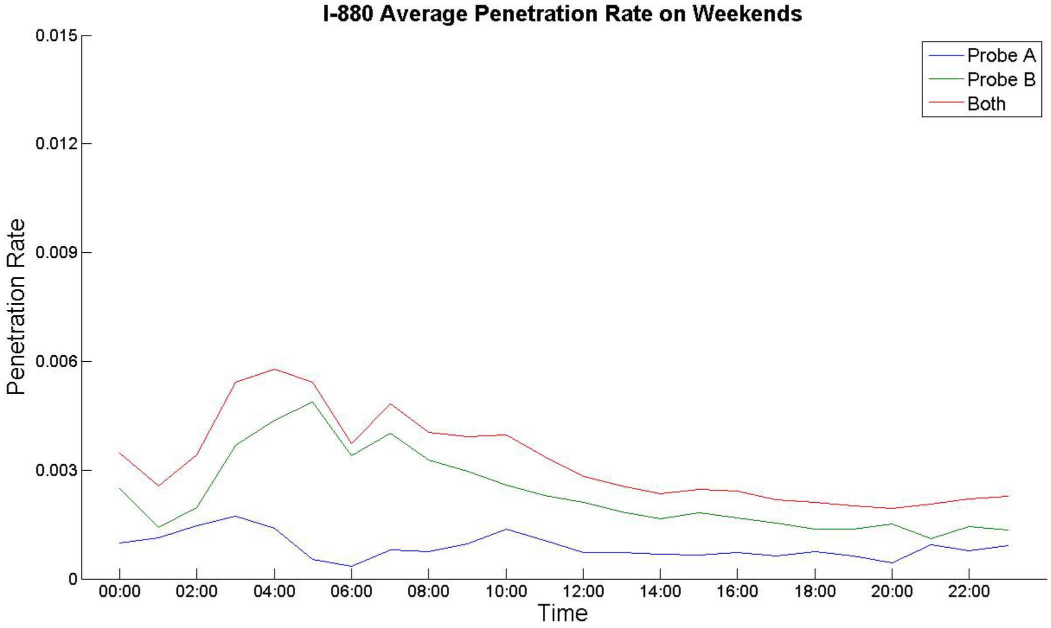


Figure 5-43: Average penetration rate on I-880 weekends

Again, we take the 85th percentile of the data rather than the average to see what a "good day" of data looks like. Probe flow here doesn't seem to increase as much relative to PeMS flow, and so we see still a 0.1–0.6% penetration rate at most times, although the spike at 4:00 am seems to show a 0.9% penetration rate.

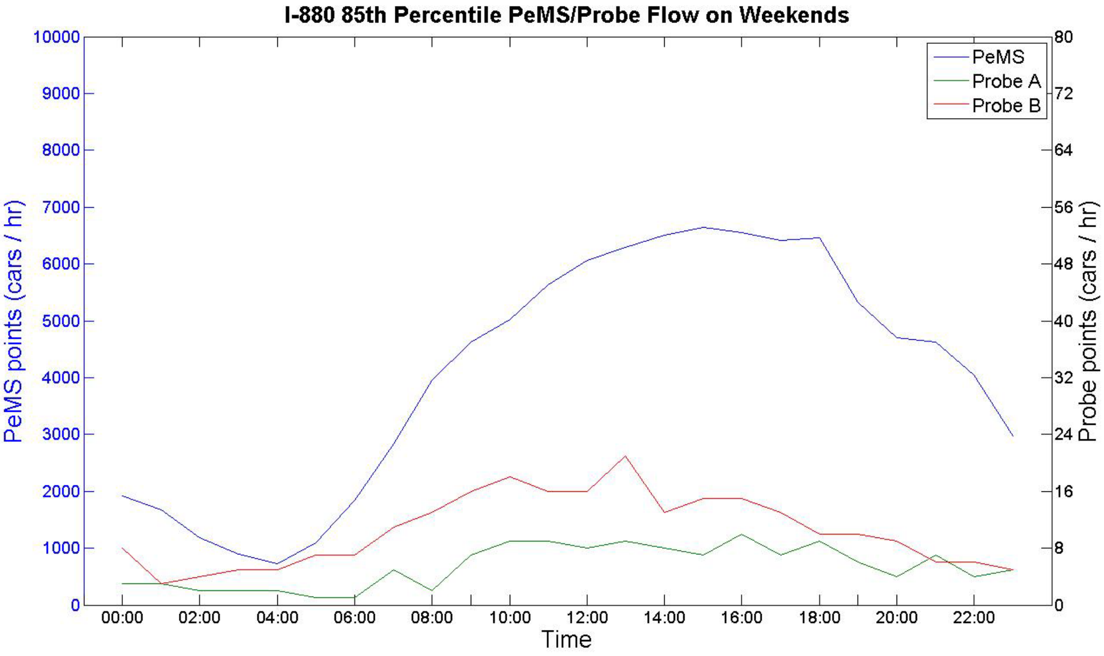


Figure 5-44: 85th percentile PeMS/probe flow on I-880 weekends

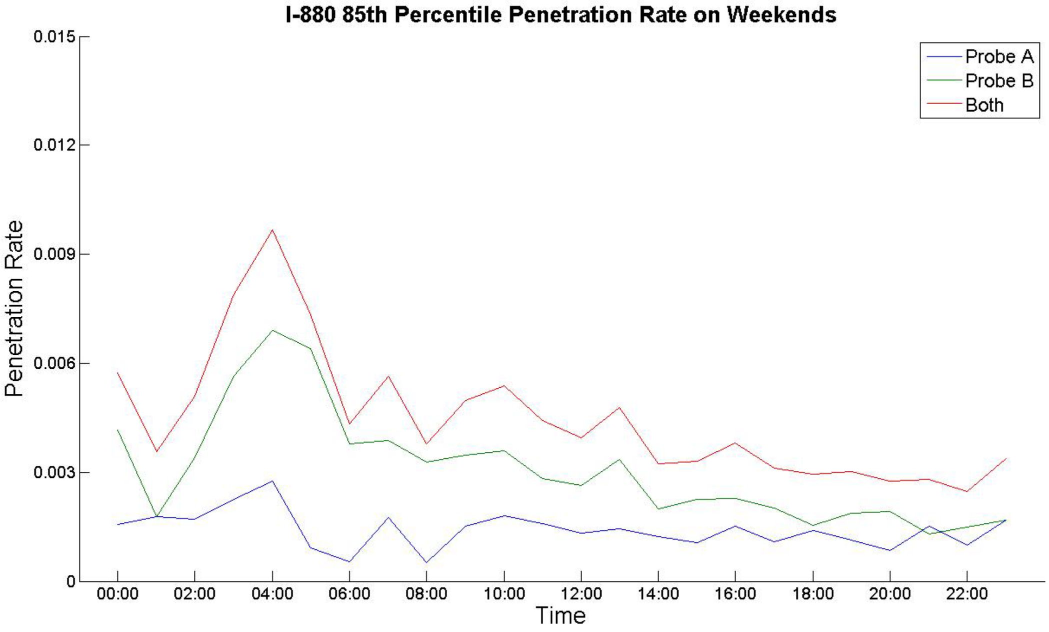


Figure 5-45: 85th percentile penetration rate on I-880 weekends

7.3.3 I-15 ONTARIO WEEKDAYS

Figure 5-46 and Figure 5-47 show the average weekday probe and PeMS flows on the I-15 Ontario site and the corresponding penetration rates. Probe B flow is lower at this site than on I-880, although unlike I-880 it seems to peak at the same time as the PeMS flow rather than slightly before. The penetration rates tend to lie between 0.3% and 0.6% consistently throughout the day.

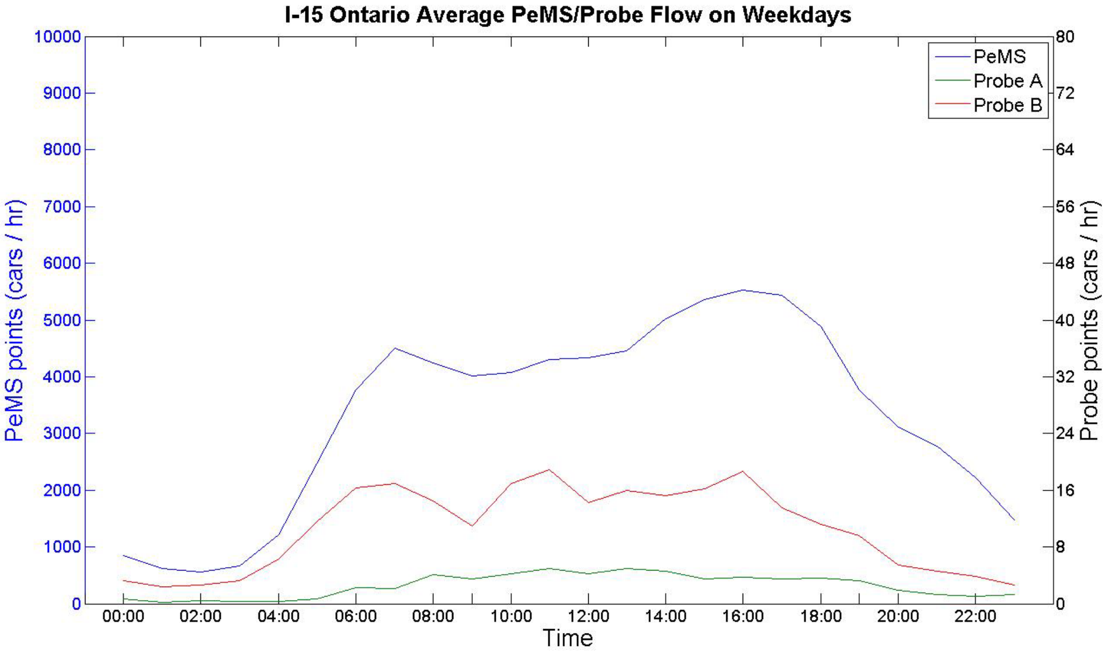


Figure 5-46: Average PeMS/probe flow on I-15 Ontario weekdays

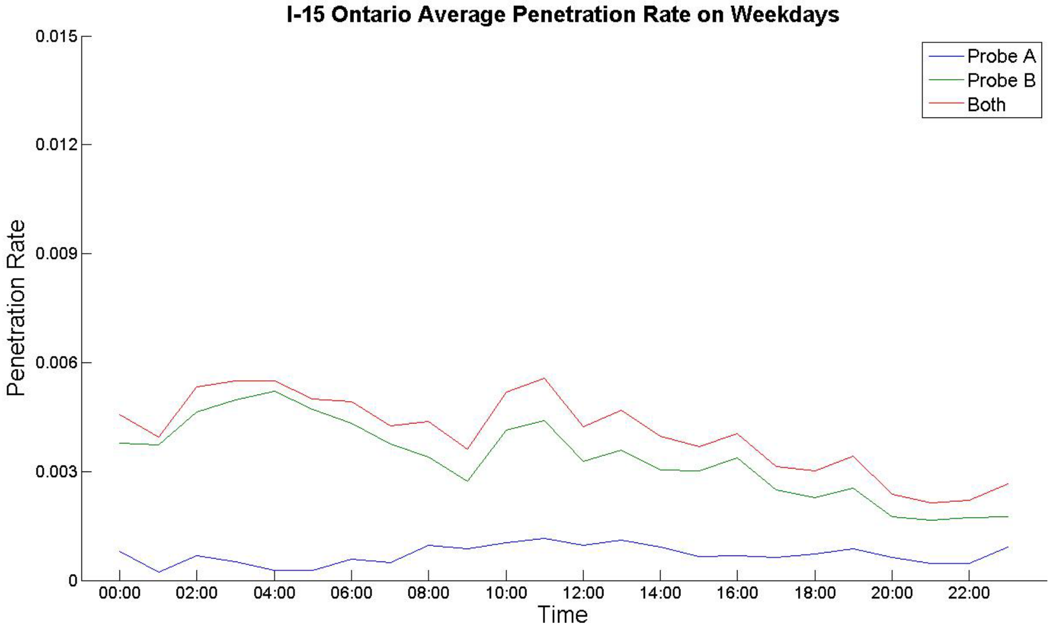


Figure 5-47: Average penetration rate on I-15 Ontario weekdays

Taking the 85th percentile of the data does not seem to change the penetration rates very much, although it is interesting to note the large peak that develops in the morning for PeMS flow. This seems to indicate that some mornings have much worse rush hour traffic than others at this site, and may indicate that the probe flow does not increase accordingly when this happens, since probe flow does not show a corresponding peak.

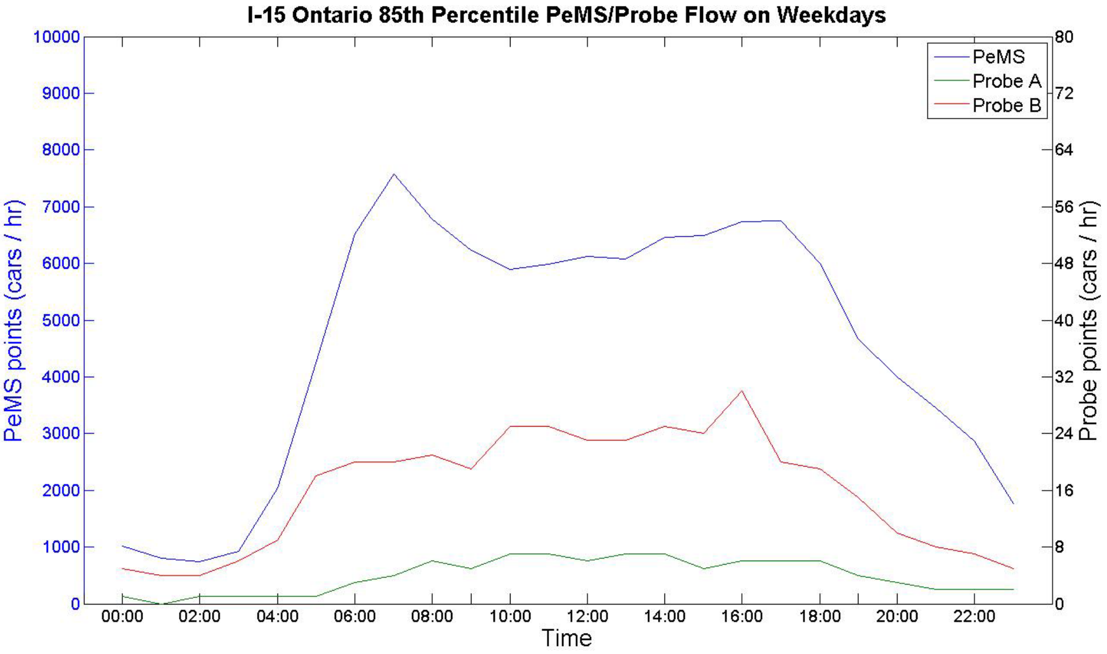


Figure 5-48: 85th percentile PeMS/probe flow on I-15 Ontario weekdays

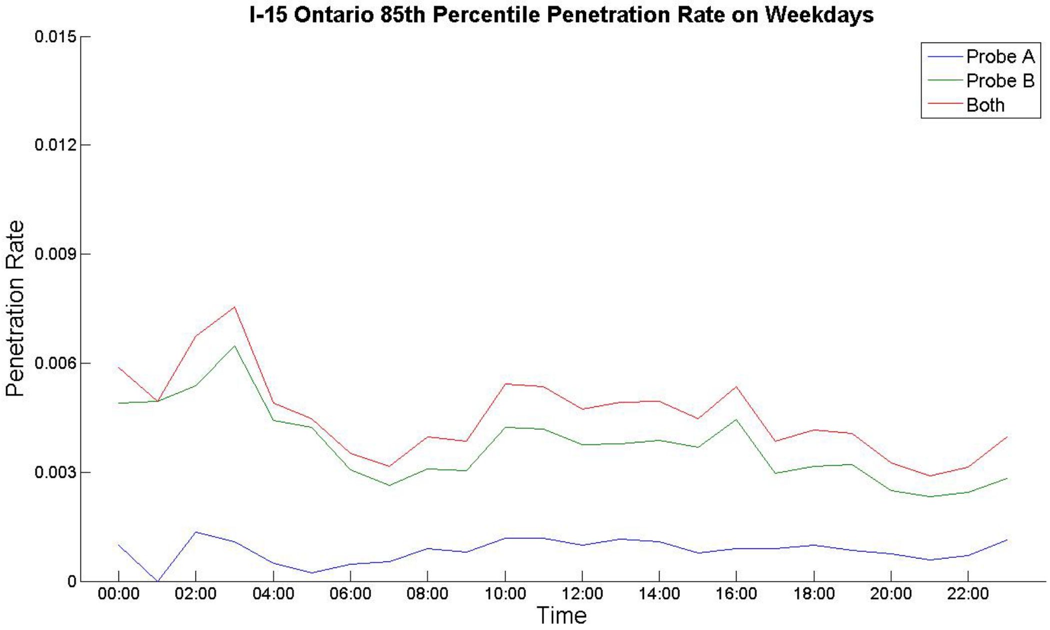


Figure 5-49: 85th percentile penetration rate on I-15 Ontario weekdays

7.3.4 I-15 ONTARIO WEEKENDS

Figure 5-50 and Figure 5-51 show flows and penetration rates for weekend days on the I-15 Ontario site. Again, when compared to I-880, probe B yields much lower flows at this site. Thus, the penetration rates consistently stay between 0.1% and 0.3% throughout the day.

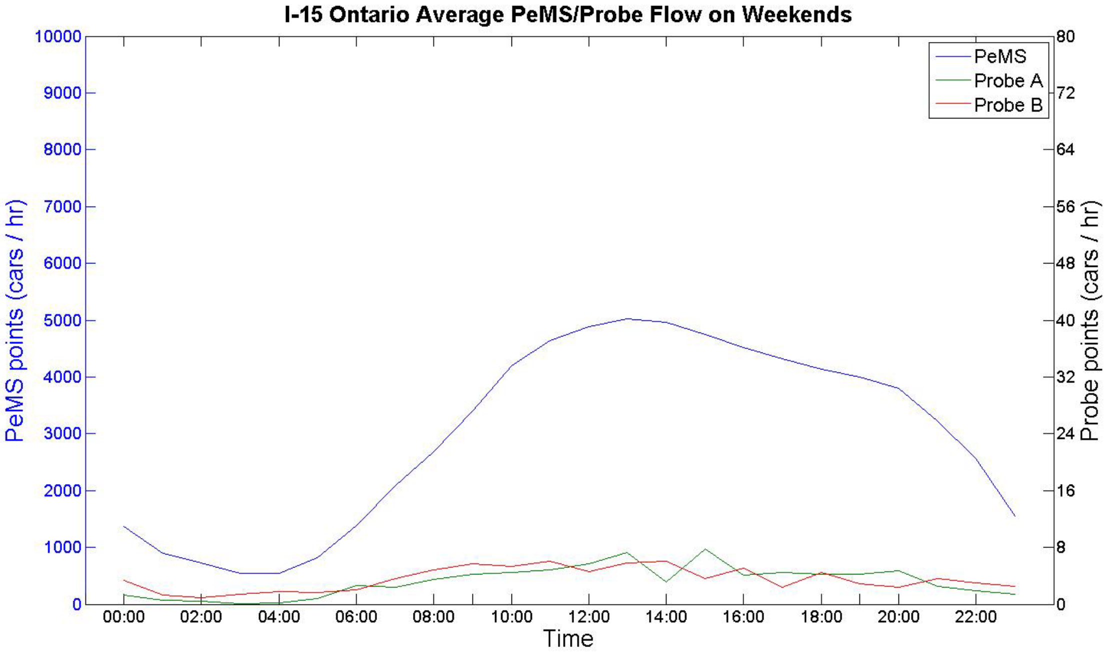


Figure 5-50: Average PeMS/probe flow on I-15 Ontario weekends

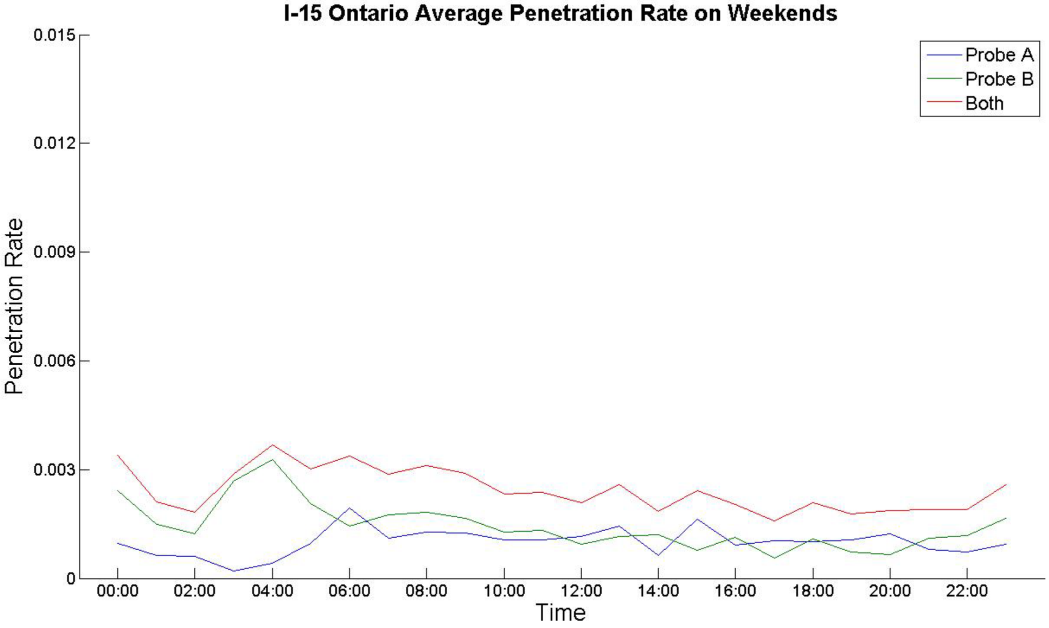


Figure 5-51: Average penetration rate on I-15 Ontario weekends

These plots show the 85th percentile of flows and penetration rates on weekends on I-15 Ontario. Again, there is not very much difference here compared to the average penetration rates, and we see consistent penetration rates between 0.1% and 0.5%.

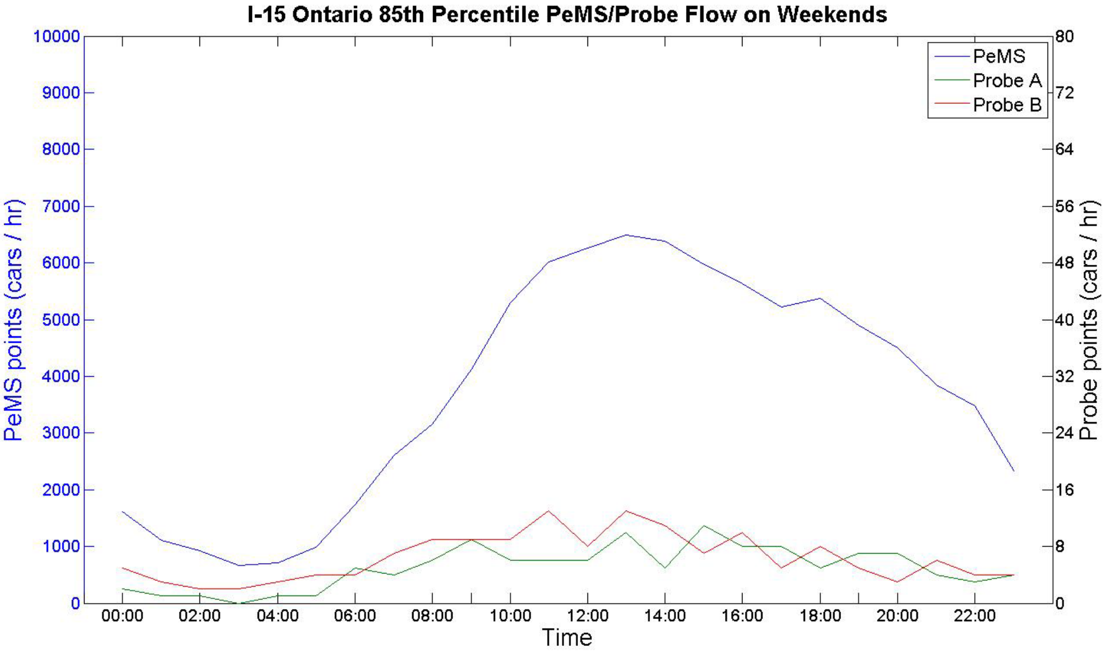


Figure 5-52: 85th percentile PeMS/probe flow on I-15 Ontario weekends

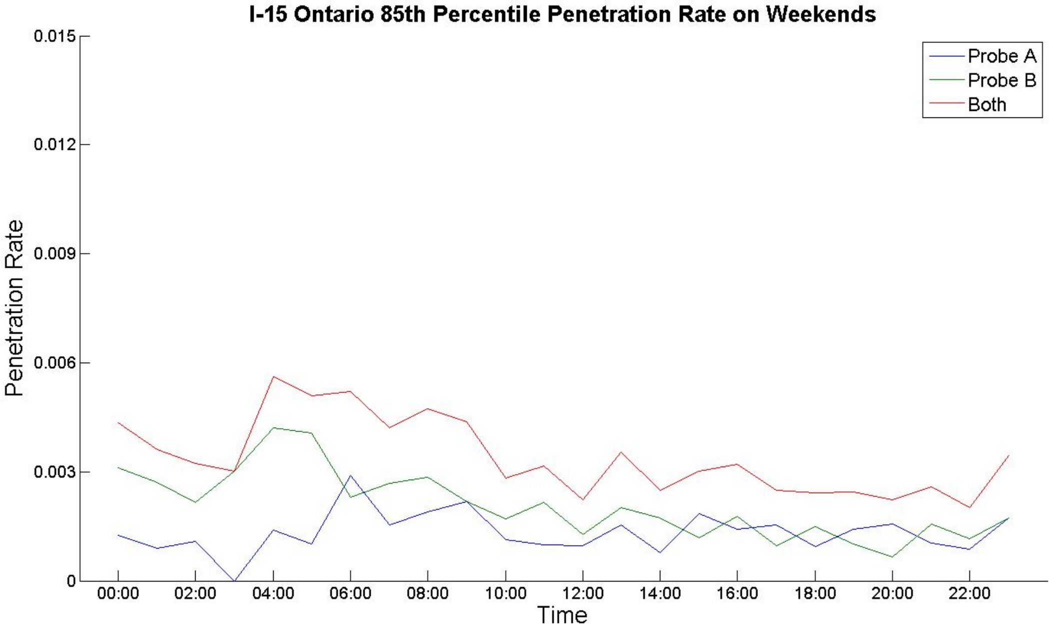


Figure 5-53: 85th percentile penetration rate on I-15 Ontario weekends

7.3.5 OVERALL PERCENTILE FLOWS

In addition to plotting flows and penetration rates for weekdays and weekends separately, we also calculated the 85th percentile flows over the whole time period by taking all data (both weekday and weekend) from all three PeMS stations used across the entire time range and selecting the 85th percentile of each. The results are as follows:

I-880 85th Percentile Flows	**I-15 Ontario 85th Percentile Flows**
PeMS: 5,931 cars per hour Probe A: 7 cars per hour Probe B: 35 cars per hour	PeMS: 5,688 cars per hour Probe A: 5 cars per hour Probe B: 18 cars per hour
Penetration rate: (7+35)/5931 = 0.7%	**Penetration rate:** (5+18)/5688 = 0.4%

These results also reinforce the plot results that show probe B flows were much higher on I-880 than on I-15 Ontario.

7.3.6 COMPARISON TO BLUETOOTH

For comparison, we also calculated the Bluetooth flows and penetration rates, focusing on I-880. To do so, we took the same three centrally located PeMS stations used for analyzing the probe flows, and looked at the Bluetooth detectors closest to those to get data for Bluetooth flows. We collected all the data and computed both the average and 85th percentile flows and penetration rates over weekdays, weekends, and overall.

	I-880 Average Percentile Flows	**I-880 85th Percentile Flows**
Weekdays	PeMS: 3,895 cars per hour Bluetooth: 255 cars per hour **Penetration rate:** 255/3895 = 6.56%	PeMS: 5,931 cars per hour Bluetooth: 388 cars per hour **Penetration rate:** 388/5931 = 6.53%
Weekends	PeMS: 3,445 cars per hour Bluetooth: 230 cars per hour **Penetration rate:** 230/3445 = 6.67%	PeMS: 5,885 cars per hour Bluetooth: 383 cars per hour **Penetration rate:** 383/5885 = 6.51%
Overall	PeMS: 3,754 cars per hour Bluetooth: 247 cars per hour **Penetration rate:** 247/3754 = 6.59%	PeMS: 5,931 cars per hour Bluetooth: 387 cars per hour **Penetration rate:** 387/5931 = 6.53%

8 RESULTS FOR I-15 ONTARIO

Results similar to those for I-880 were achieved for the portion of I-15 Ontario in the study.

8.1 BLUETOOTH TRAVEL TIME OVER TWO WEEKS

Figure 5-54 and Figure 5-55 summarize two weeks of Bluetooth travel time data collected on I-15 Ontario. (For a description of Bluetooth travel time calculations, see section 3.1.)

- Areas of congestion are numbered and identified within black bounding boxes in the figures.
- Metrics that are referred to as “global” in section 8.7 are calculated with respect to those numbered areas only.

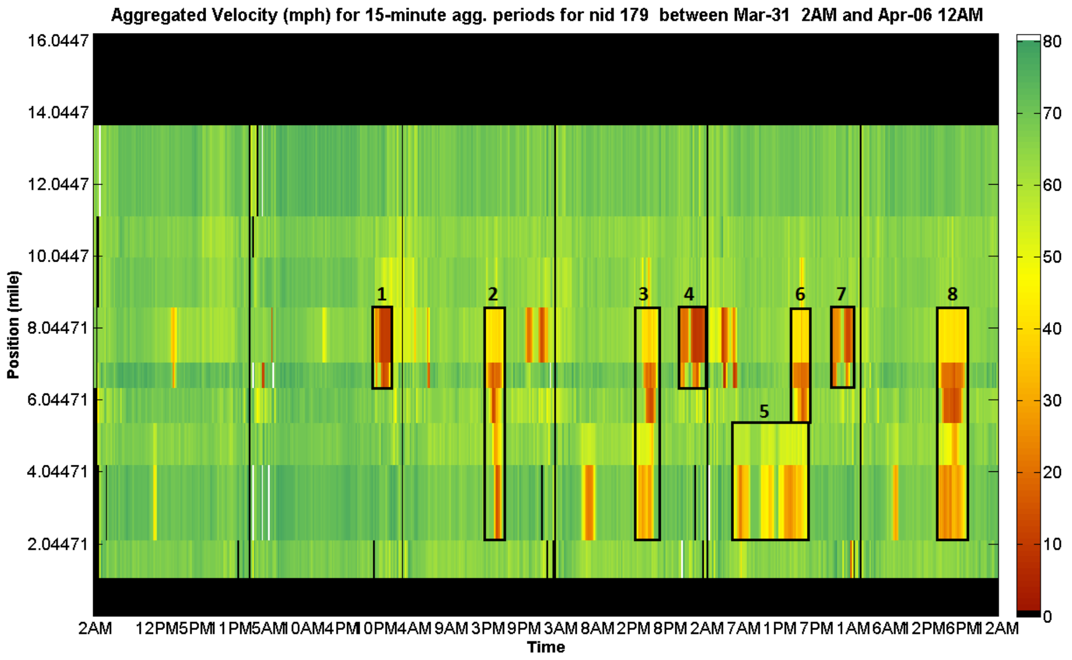


Figure 5-54: Ground truth traffic conditions as derived from Bluetooth travel times and presented as average velocities. The data is for one week from Saturday, March 31 to Friday, April 6. Black horizontal bands at the top and bottom of the graph (absence of data) indicate that the highway segment used in the model (the mathematical representation of the road network) is slightly longer than the actual length of the Bluetooth deployment.

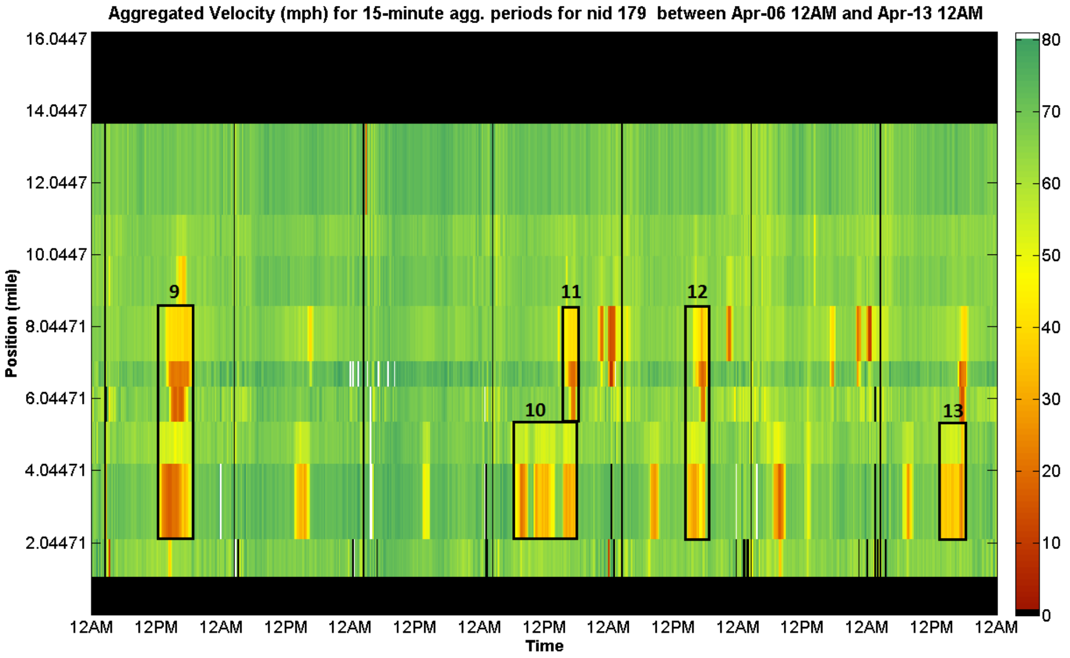


Figure 5-55: Ground truth traffic conditions as derived from Bluetooth travel times and presented as average velocities. The data is for one week from Friday, April 6 to Friday, April 13.

In Figure 5-54 and Figure 5-55, the patterns of congestion are localized to a small portion of the study site (between miles 2 and 9, between BT station 2 and BT station 7). From this observation, we can already guess that the influence of loop detectors after mile 9 will be less in our way of computing the global metrics. Similarly, probe data points after mile 9 or before mile 2 will not have much importance. This is useful to keep in mind when looking at the results presented in the next sections.

8.2 CTM+ENKF VELOCITY MAPS

Visual results of data assimilation. The data fusion engine (the velocity estimation algorithm) takes as input the filtered data from loops and probes and reconstructs the traffic state over the study site, using the well-known CTM and EnKF as described in sections 2.3 and 2.4. The spatial resolution varies slightly depending on road geometry but is approximately 200 meters. The time resolution is 30 seconds.

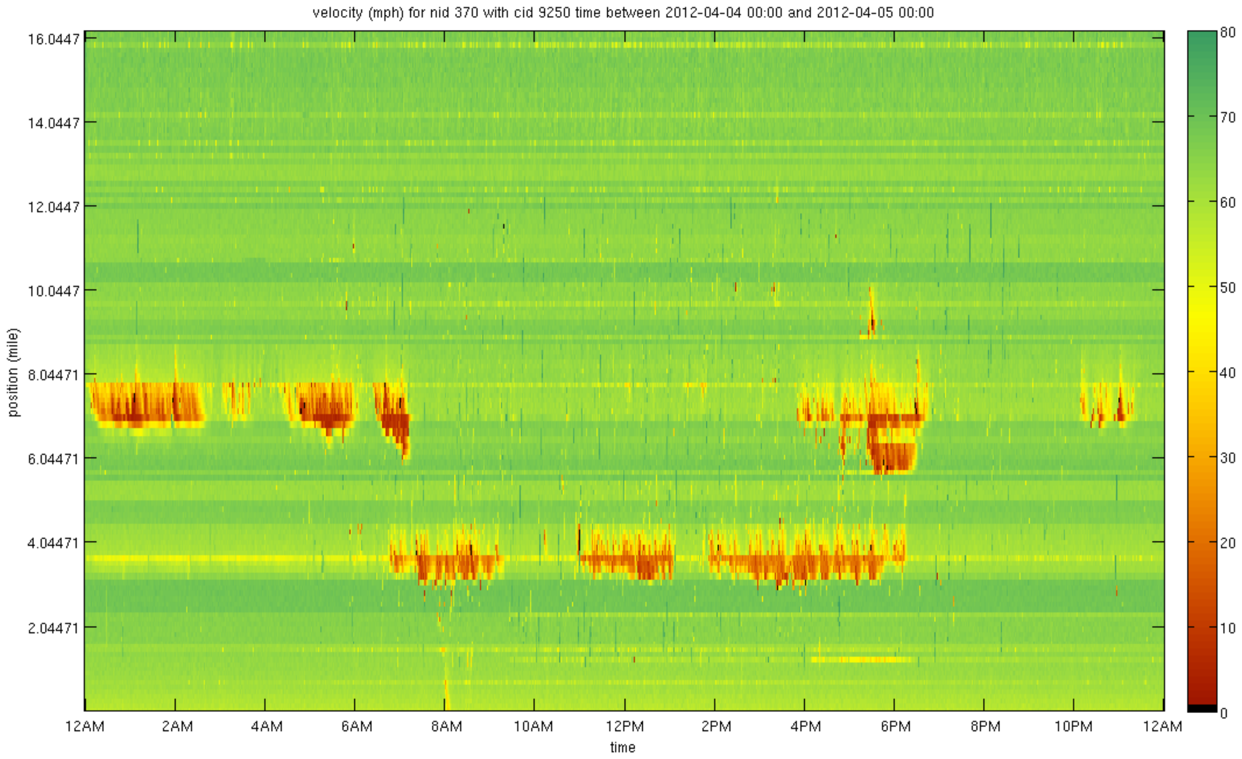


Figure 5-56: Reconstruction of traffic velocity state on April 4 using both loop data and probe data.

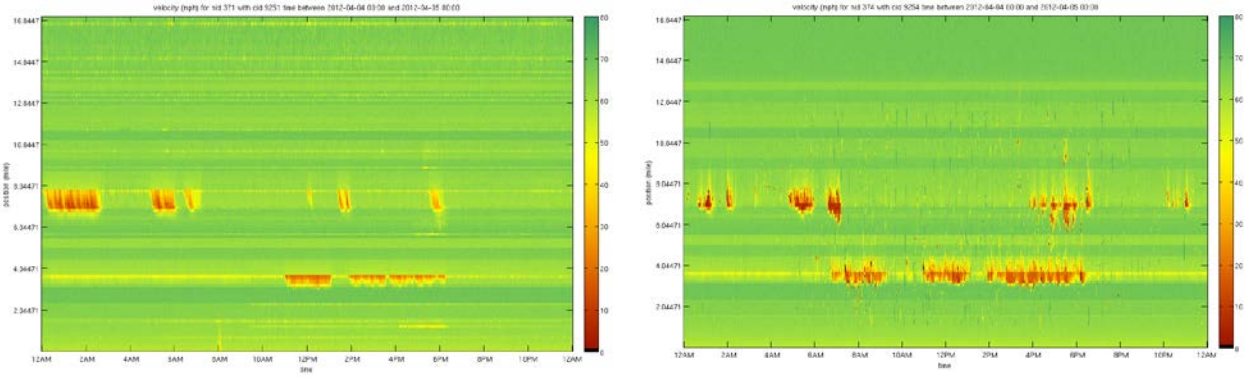


Figure 5-57: Reconstruction of traffic velocity state using loop data only (left) and probe data only (right). These results cover the same regions of space-time along I-15 Ontario as shown in Figure 5-56.

8.3 COMPARISON WITH BLUETOOTH AGGREGATED VELOCITY MAPS

As described in section 3, the quality of the traffic state reconstruction is assessed by comparing it to sampled travel times from Bluetooth detectors. Although presented here as velocities for clarity, the raw measurements being compared are travel times and not velocities.

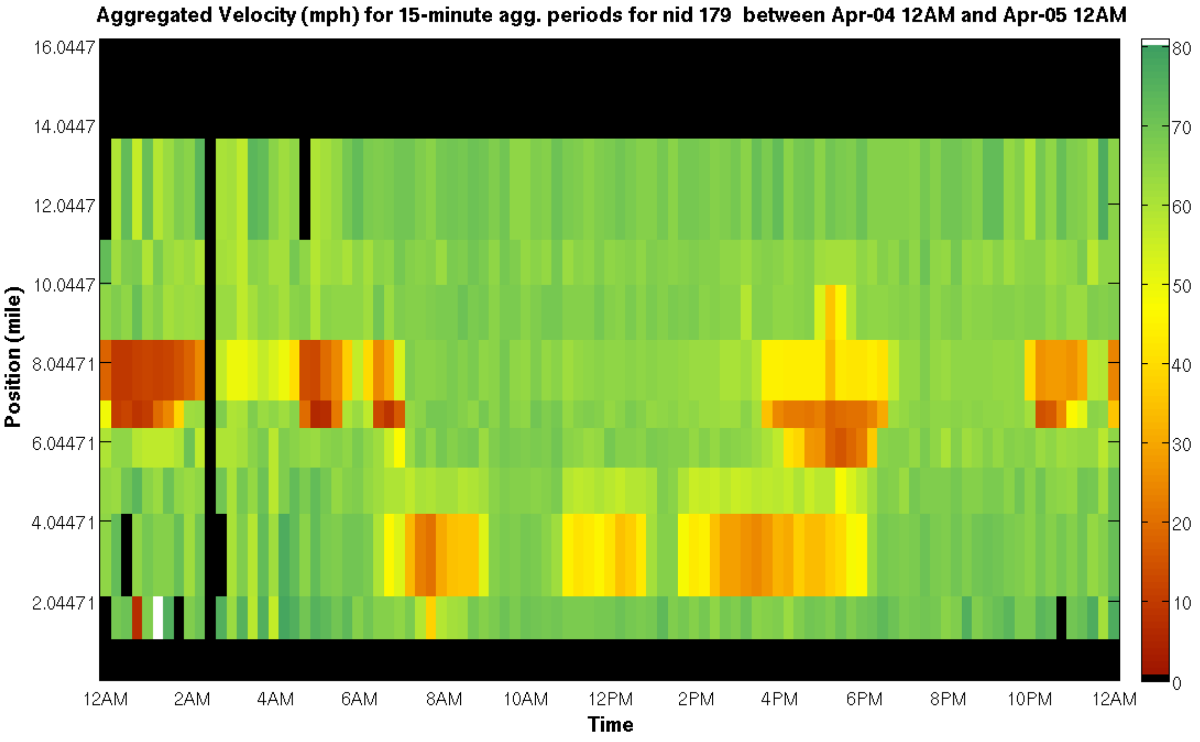


Figure 5-58: Average velocities derived from Bluetooth-detected travel times on April 4 measured over 15-minute intervals along each of the routes between successive Bluetooth sensors.

In order to have an apples-to-apples comparison, the estimated velocities from the data fusion engine (as shown in section 8.2) must be used to generate vehicle trajectories along the same routes between successive Bluetooth detectors as shown in Figure 5-58. These generated trajectories are then used to calculate a set of travel times, τ'_{jk}. These travel times are then averaged in a way consistent with the processing of Bluetooth data.

$$T_{hwy}(i,j) = \frac{1}{\sum_{iT \le k < (i+1)T} 1} \sum_{iT \le k < (i+1)T} \tau'_{jk}$$

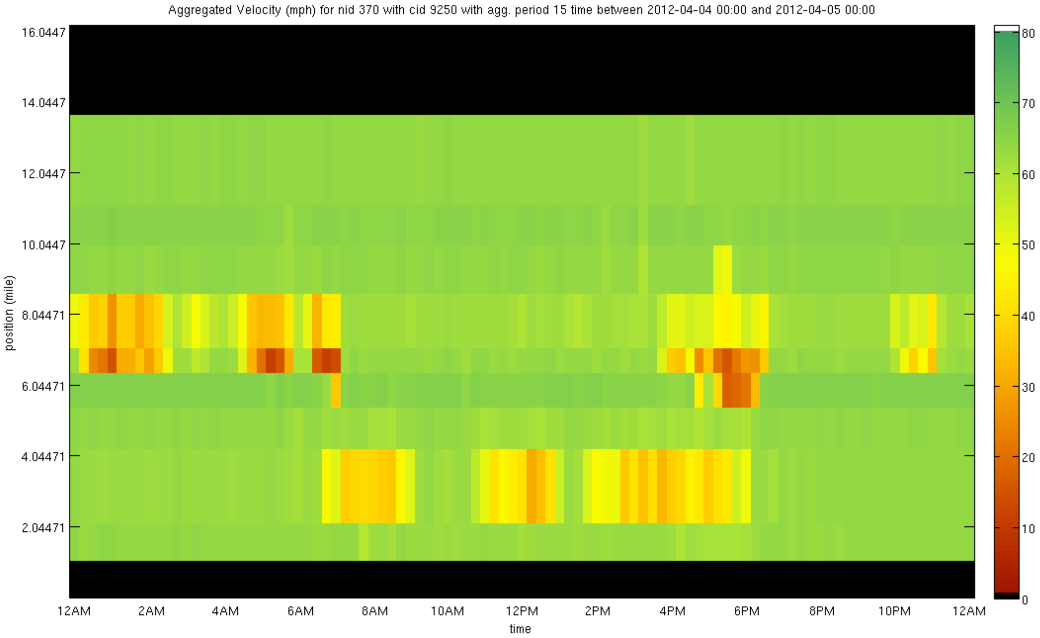


Figure 5-59: Average velocities on April 4 calculated from the reconstruction of vehicle trajectories using both loop and probe data. The velocities are calculated over 15-minute intervals along each of the routes between successive Bluetooth detectors.

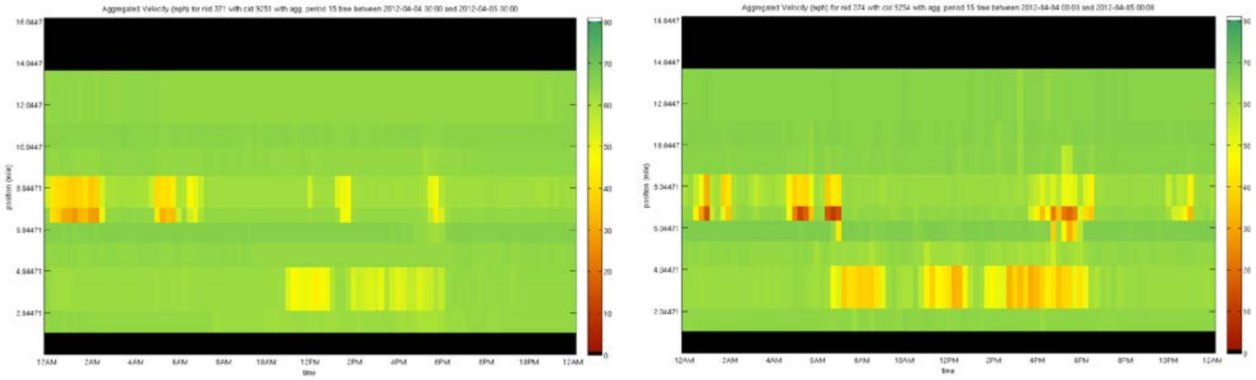


Figure 5-60: Average velocities on April 4 calculated from the reconstruction of vehicle trajectories using only loop data (left) and only probe data (right). The velocities are calculated over 15-minute intervals along each of the routes between successive Bluetooth detectors.

8.4 LOOP STATIONS ON I-15 ONTARIO

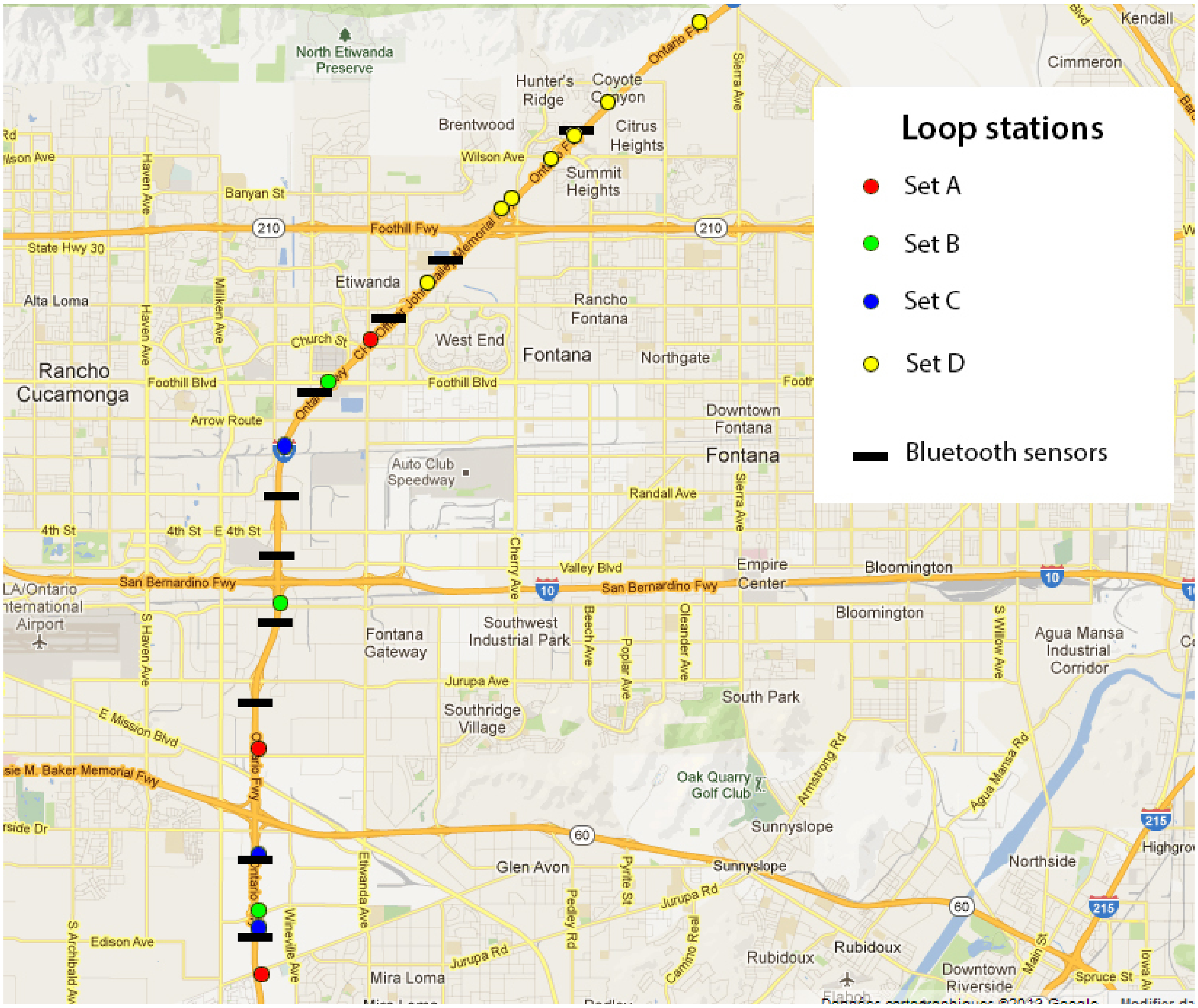


Figure 5-61: Loop stations used in the study of I-15 Ontario

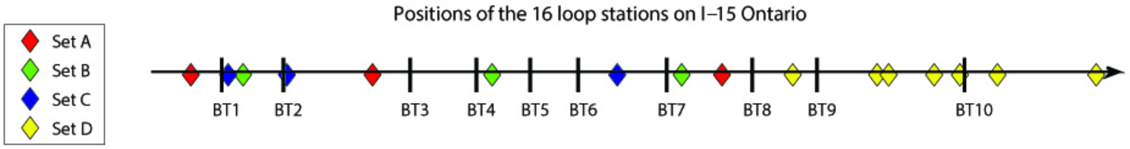


Figure 5-62: Loop stations used in the study of I-15, showing the same information as Figure 5-61, without the map

The freeway section of interest had 16 loop stations. To include different numbers of loop stations in the fusion calculations, we divided them into four sets, depending on their order of inclusion in the data fusion engine.

To include this:	**We used this:**
Three loop stations	Set A
Six loop stations	Sets A + B
Nine loop stations	Sets A + B + C
All the loop stations	Sets A + B + C + D

As with I-880 (described in section 6.1), a heuristic method was used to group loop stations into appropriate sets, but the purpose was not the same as for the I-880 network. For I-15 Ontario, the first 9 loop stations were identified as a priori more important than the last 7. As can be seen in Figure 5-62, the last 7 stations (Set D) are located after the 8th Bluetooth station, whereas we saw in section 8.1 that the patterns of congestion measured by Bluetooth sensors were localized between the 2nd and the 7th Bluetooth station. Therefore, we decided to add loop stations 1–9 to the model first, in three steps (three sets), and then add the remaining stations in the last set. The incremental addition of loop stations is shown in Figure 5-63:

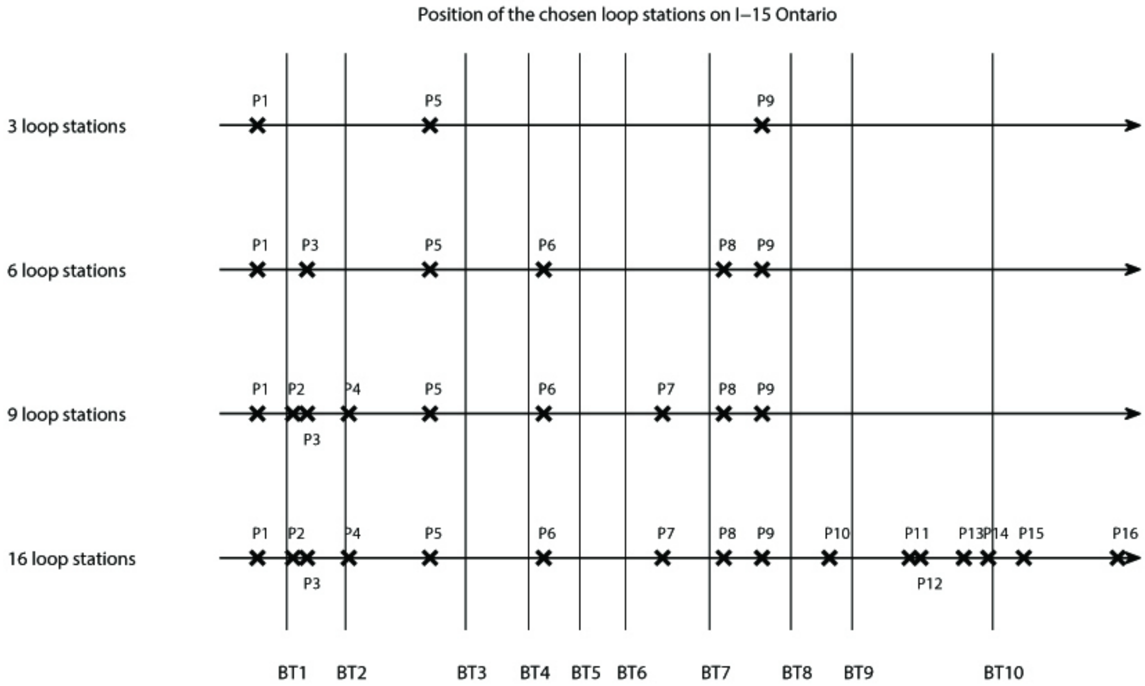


Figure 5-63: Different sets of loop stations on I-15 used in the data fusion engine. From top to bottom, loop stations included when using three, six, nine, and all of the loop stations available and active during the period of the experiment.

8.5 METRIC VS. NUMBER OF UNIQUE PROBE DEVICES FOR FIXED FRACTIONS OF LOOPS

MAPE. Figure 5-64 shows that probe data can be used to improve or to supplement loop detector data for the purpose of estimating travel times. For each line series, the number of loop detector stations was fixed. The amount of probe data was adjusted using the method described in section 7.1.

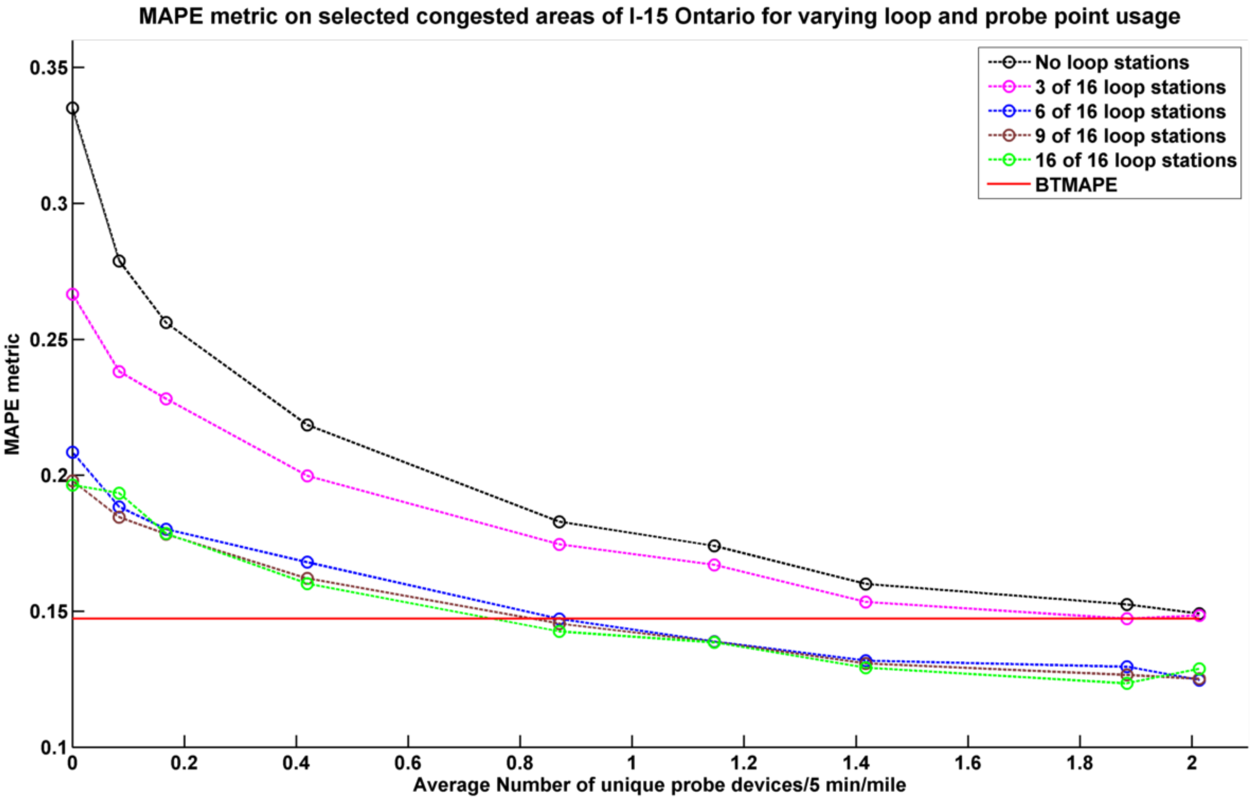


Figure 5-64: Global travel time MAPE plotted against the average number of unique probe devices used per five minutes per mile. Each dashed line corresponds to the performance achievable when fusing probe data with that from a different fraction of available loop detector stations. The red line corresponds to the “noise floor” of measured travel time variability.

For example, when using all the loop data but no probes, the performance metric is 0.2. This same level of performance can also be achieved by using 3 of the 16 loop stations and having probe data corresponding to at least 0.4 unique probe devices per five minutes per mile. Using all the loop data, the metric is further reduced to less than 0.15 (the “noise floor”) by adding data from an average of 0.8 unique probe devices per five minutes per mile.

It can also clearly be seen that the metric improves by adding 3 loop stations and then significantly more by adding 6 loop stations. Beyond that, however, performance is virtually the same whether using 6, 9, or 16 loop stations.

PMATE. Not surprisingly, Figure 5-65 yields the same conclusions as Figure 5-64, but this time using the PMATE metric. When all loop data is available but no probes, the metric is 20 seconds per mile. For the 15-mile section of freeway, it will incur in the worst case a 5-minute error in travel time prediction. Adding probes makes it possible to lower the metric to 12 seconds/mile, which results in a 3-minute error in the worst case. As with the MAPE shown in Figure 5-64, this occurs whether using 16, 9, or 6 loop stations.

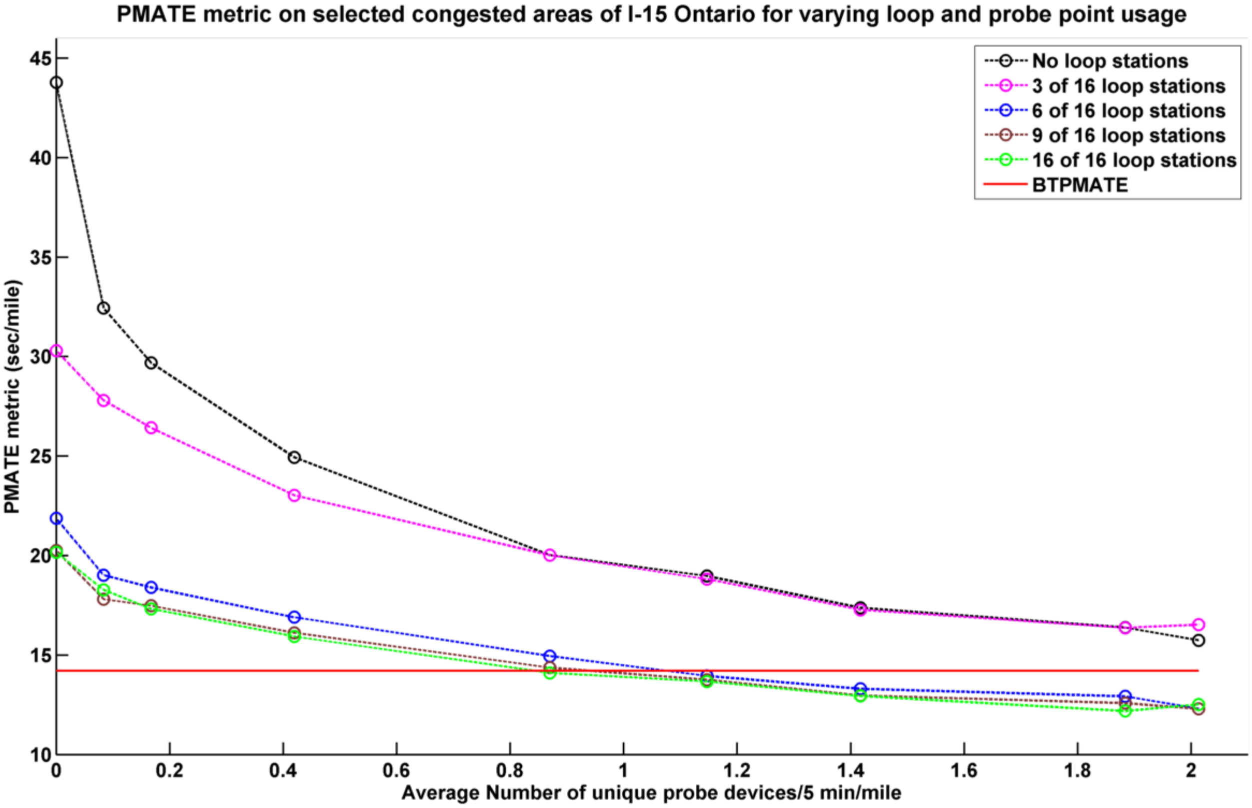


Figure 5-65: Global travel time PMATE plotted against the average number of unique probe devices used per five minutes per mile. Each dashed line corresponds to the performance achievable when fusing probe data with that from a different fraction of available loop detector stations. The red line corresponds to the "noise floor" of measured travel time variability.

CCEC. The behavior of the congestion classification metric is very instructive when fusing loop and probe data. In Figure 5-66, for example, the metric is low (0.11, or 11% difference from Bluetooth) when using all loop data but no probes, and there is little improvement when probe data is added. That same metric is attained when using probe data only (with 1.9 unique probe devices per five minutes per mile), showing a reduction in the metric from 1 (no data, 100% difference from the Bluetooth measure of congestion) to 0.11.

This is an example where using a certain set of loop stations allows a good congestion estimate even without probe data. This is because the CCEC metric is far less precise than the other metrics. However, when only 3 loop stations are used, adding probe data is very useful. With only 0.1 unique probe devices per five minutes per mile, the CCEC metric is reduced from 0.57 to 0.34. The metric is further reduced to less than 0.1 when using only 6 of 16 loop stations and 0.1 unique probe devices per five minutes per mile.

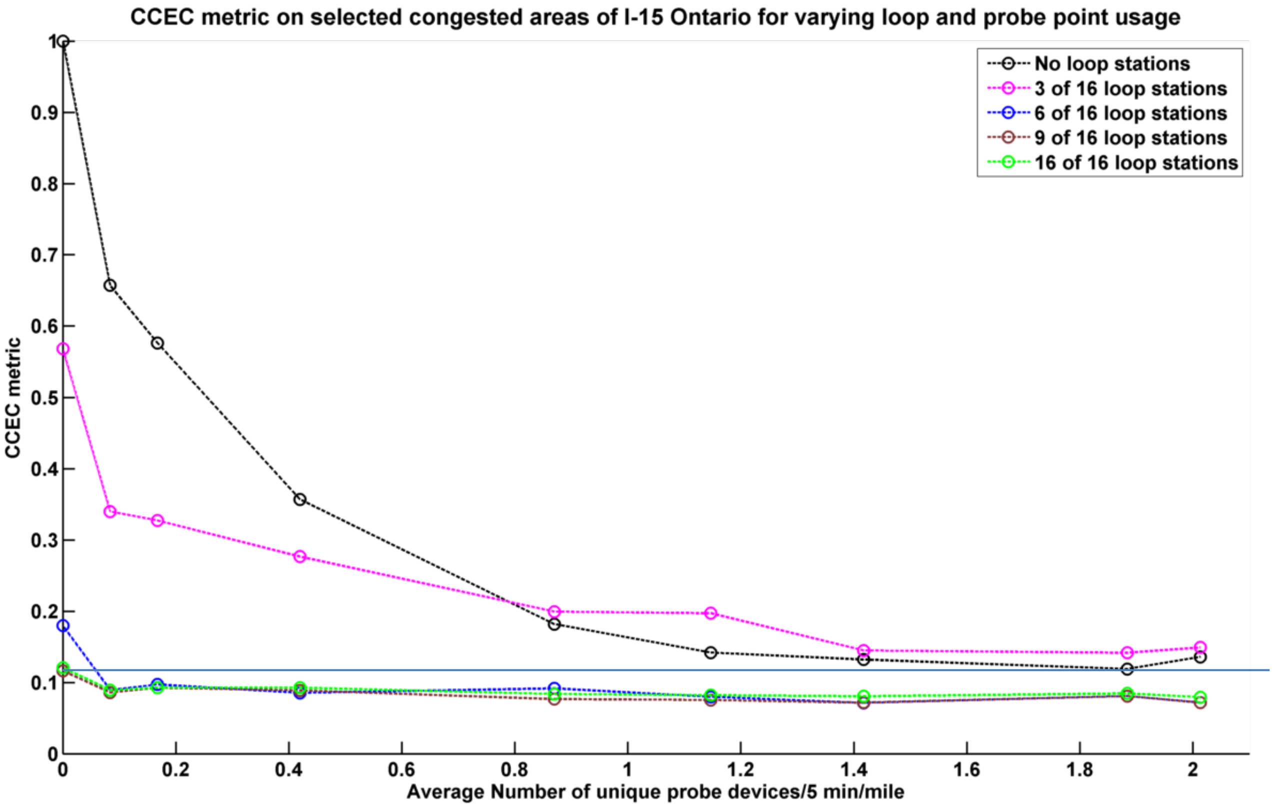


Figure 5-66: CCEC for I-15 Ontario for varying loop stations, plotted against the average number of unique probe devices used per five minutes per mile. Each dashed line corresponds to the performance achievable when fusing probe data with that from a different fraction of available loop detector stations.

8.6 DATA FUSION ISOMETRICS

As described in section 6.3 for highway I-880, the data fusion isometrics directly compare the predictive power of loop vs. probe data. For example, Figure 5-67 displays combinations of probe and loop data for I-15 Ontario that could be used to achieve a given MAPE.

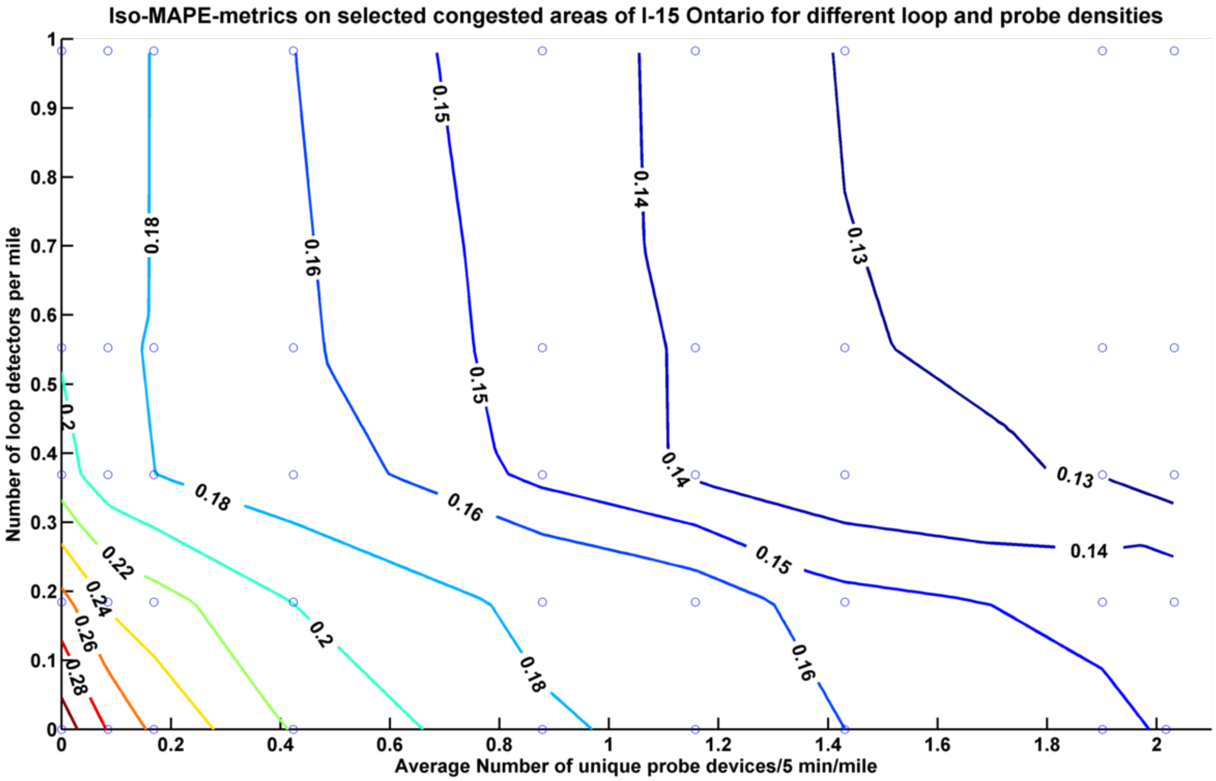


Figure 5-67: Contours of constant travel time estimation performance as a function of both number of loop detectors and average number of unique probe vehicles per five minutes per mile. The lines represent the MAPE (percent discrepancy from Bluetooth measurements), with only probe data along the x-axis and only loop data along the y-axis.

On this plot, a MAPE of 0.2 (20%) is obtained with approximately 0.5 loop stations per mile with no probe data, or slightly more than 0.6 unique probe devices per five minutes per mile with no loop data. Between those two endpoints is a spectrum of loop/probe combinations along the 0.2 contour. The vertical contour lines in the upper part of the plot suggest that little is gained by adding more than 0.5 loop stations per mile, while adding probe data shows a clear improvement in the MAPE. This shape results from the pattern of congestion that forms on I-15 Ontario.

The vertical contour lines represent the transition from loop detectors 9–16 on the freeway (described as Set D in section 8.4), located near the end of the freeway section, where almost no congestion was detected during the time period of the experiment. Since the patterns of congestion are localized to a small portion of the study site, there is little benefit, in this example, in having loops in areas where there is little congestion.

8.7 HIGH-FREQUENCY VS. LOW-FREQUENCY PROBE DATA

As described in section 7, the performance of the data fusion engine is found to be less sensitive to the raw quantity of data and more sensitive to the quality of that data. With probe data, it is better to have less frequent data from a larger number of unique vehicles than it is to have more frequent data from a smaller number of unique vehicles.

Two categories of probe data were used for the analysis that follows:

- Type A probe data was taken from a relatively smaller population of vehicles that supplied data points at a frequency greater than one point every 10 seconds ("high-frequency" data).
- Type B probe data was taken from a relatively larger population of vehicles that provided data relatively rarely—often much less than once per minute ("low-frequency" data).

The data types were selected using the methods described in section 7.1. Figure 5-68 through Figure 5-71 show the relationship between MAPE performance on one hand and the number of unique probe devices or probe points/mile on the other hand, for probe type A and B separately and when combined. The results are similar to those shown for I-880 in section 7.1.

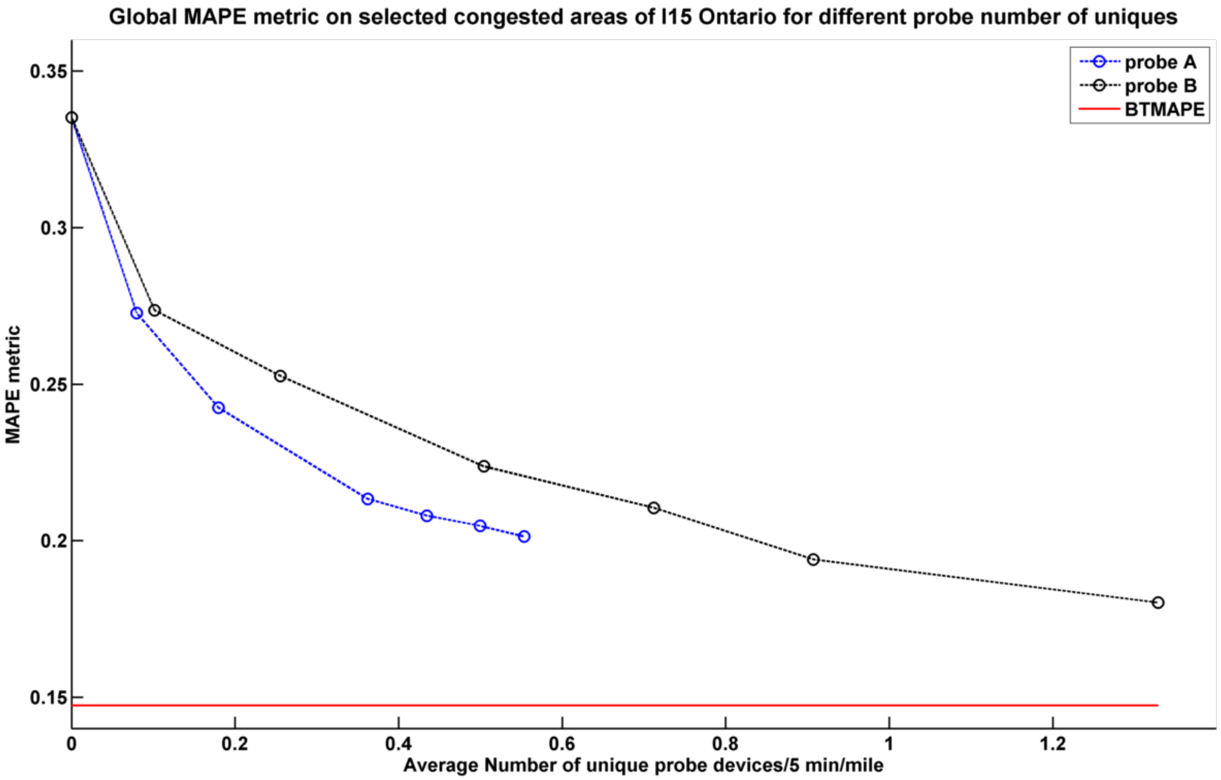


Figure 5-68: Relationship between MAPE performance and the number of unique probes for type A and B data (without PeMS data). The red line corresponds to the "noise floor" of measured travel time variability.

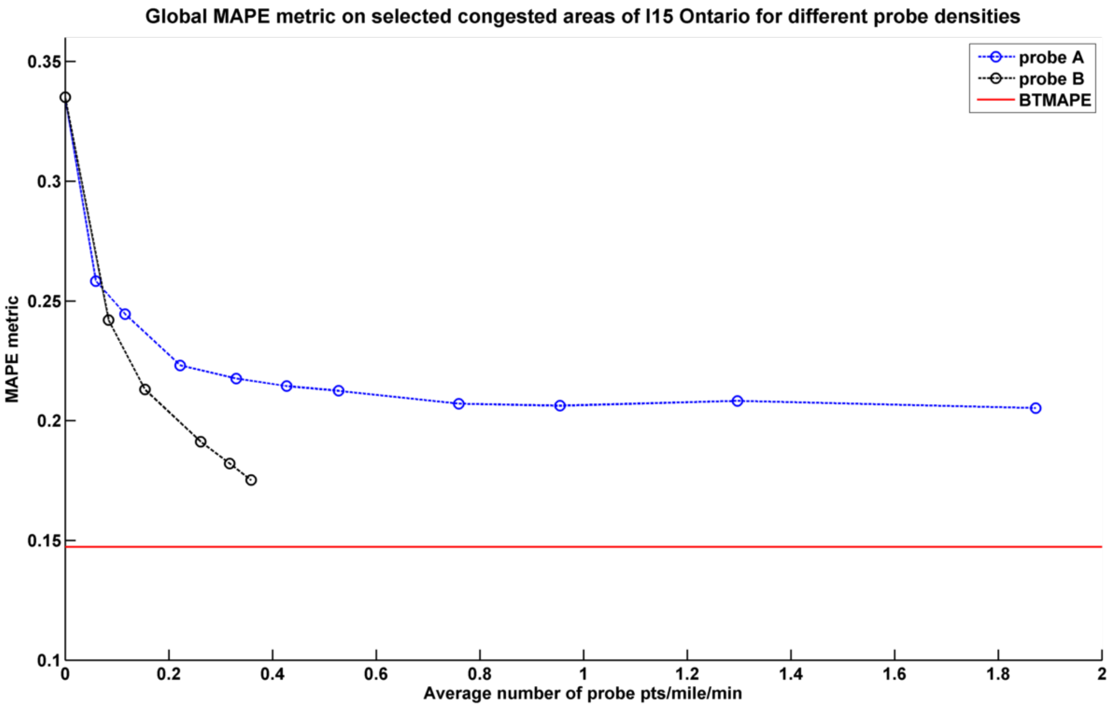


Figure 5-69: Relationship between MAPE performance and the number of probe data points for type A and B data (without PeMS data). The red line corresponds to the "noise floor" of measured travel time variability.

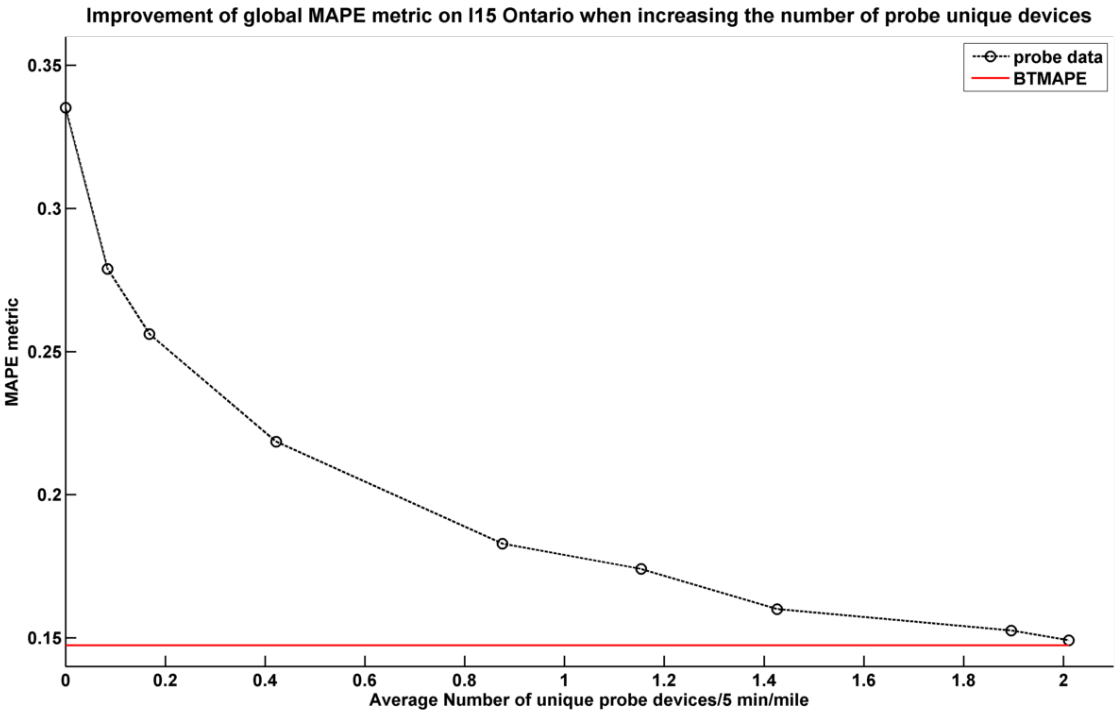


Figure 5-70: Relationship between MAPE performance and the total number of unique probes for type A and B data combined (without PeMS data). The red line corresponds to the "noise floor" of measured travel time variability.

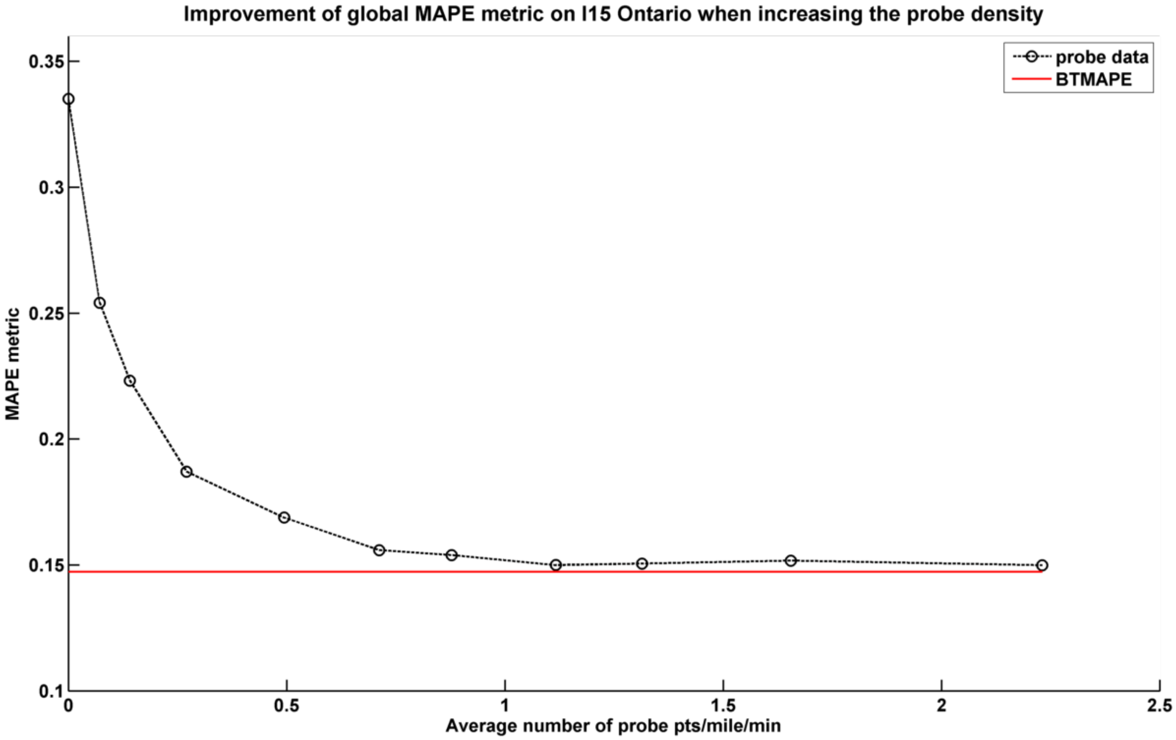


Figure 5-71: Relationship between MAPE performance and the total number of probe data points for type A and B data combined (without PeMS data). The red line corresponds to the "noise floor" of measured travel time variability.

9 RESULTS FOR I-15 VICTORVILLE

Information gathered from I-15 in Victorville southbound yielded a smaller amount of data of interest to the study than that from I-880 or I-15 Ontario. The analysis for that stretch of highway is thus based on a more limited data set.

In addition, the Victorville site was unique among selected test sites in that no loop stations were active at the time of the experiment. Therefore, it was not possible to study the effects of fusing loop data with probe data. We instead examined the ability to make estimations using probe data alone.

9.1 BLUETOOTH TRAVEL TIME

During the two weeks of Bluetooth deployment, three periods of congestion were detected over two days, as shown in the numbered bounding boxes in Figure 5-72:

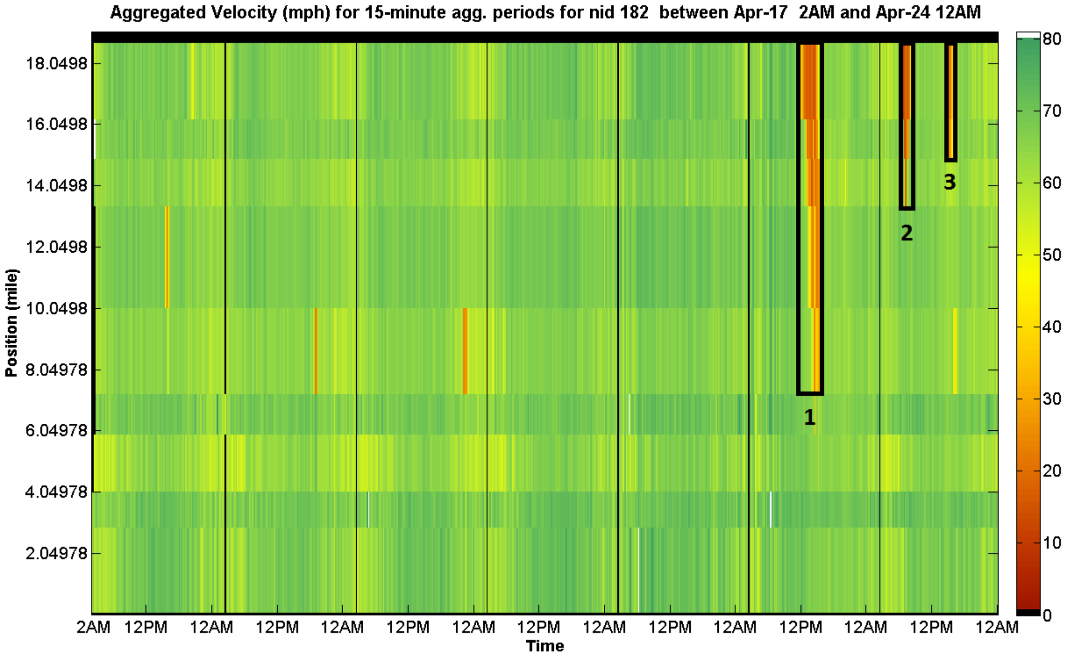


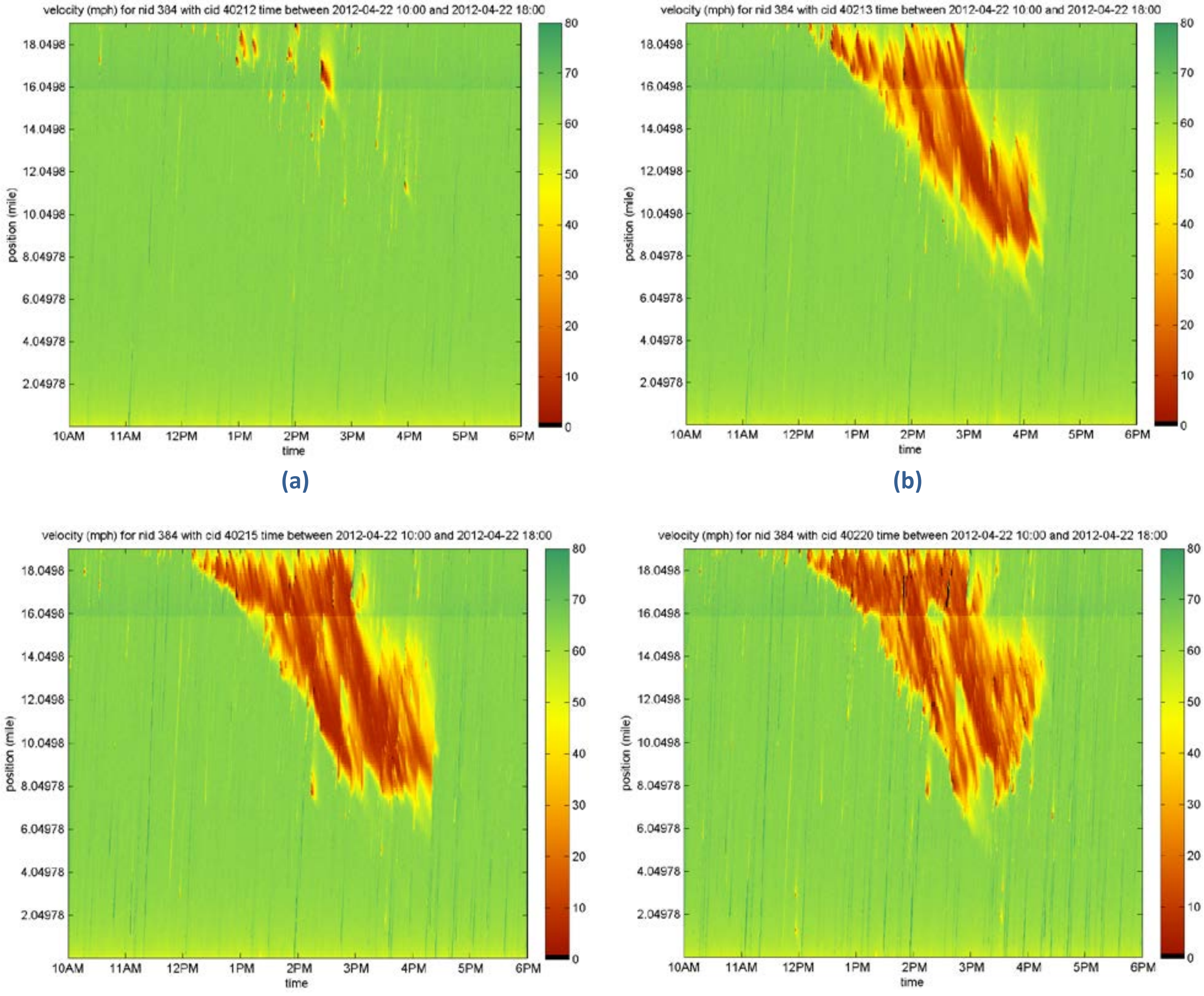
Figure 5-72: Ground truth traffic conditions as derived from Bluetooth travel times and presented as average velocities. The data is for one week from Tuesday, April 17 to Tuesday, April 24.

The "global" metrics in section 9.4 are calculated with respect to those numbered areas only.

9.2 CTM+ENKF VELOCITY MAPS

Visual results of data assimilation. For this site, the data fusion engine (the velocity estimation algorithm) takes as input the filtered data from probes only and reconstructs the traffic state using the well-known CTM and EnKF as described in sections 2.3 and 2.4. The spatial resolution varies slightly depending on road geometry but is approximately 200 meters. The time resolution is 30 seconds.

Figure 5-73 focuses on the period of congestion highlighted in box number 1 of Figure 5-72. It illustrates how the visual representation of the estimate changes as more probe data is added.



(a) (b)

(c) (d)

Figure 5-73: Reconstruction of traffic velocity state on April 22 using probe data

9.3 COMPARISON WITH BLUETOOTH AGGREGATED VELOCITY MAPS

As described in section 3, the quality of the traffic state reconstruction is assessed by comparing it to sampled travel times from Bluetooth detectors. Although presented here as velocities for clarity, the raw measurements being compared are travel times and not velocities.

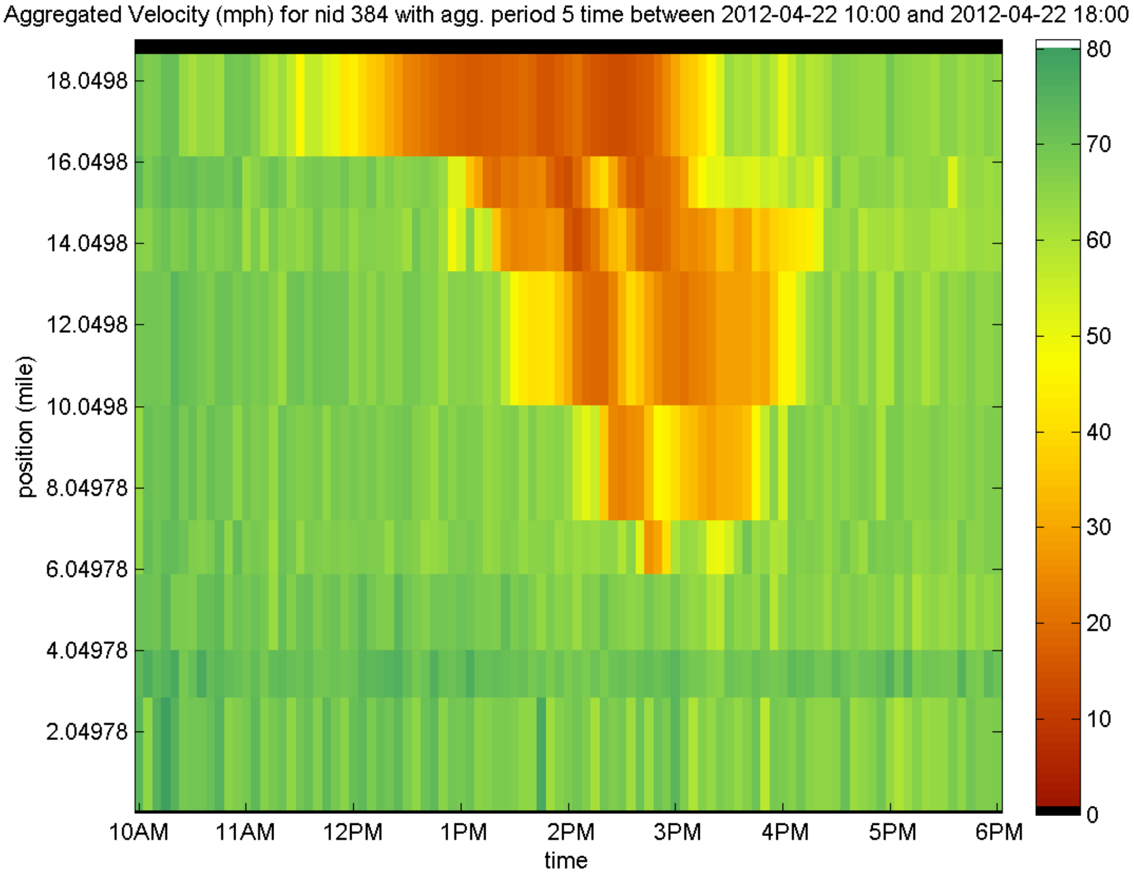


Figure 5-74: Bluetooth velocity map showing average velocities as derived from Bluetooth travel times on April 22 measured over 5-minute intervals along each of nine routes between successive Bluetooth detectors.

In order to have an apples-to-apples comparison, the estimated velocities from the data fusion engine, as shown in Figure 5-73(d), must be used to generate vehicle trajectories along the same routes between successive Bluetooth detectors as shown in Figure 5-74. These generated trajectories are then used to calculate a set of travel times, τ'_{jk}. These travel times are then averaged in a way consistent with the processing of Bluetooth data.

$$T_{hwy}(i,j) = \frac{1}{\sum_{iT \le k < (i+1)T} 1} \sum_{iT \le k < (i+1)T} \tau'_{jk}$$

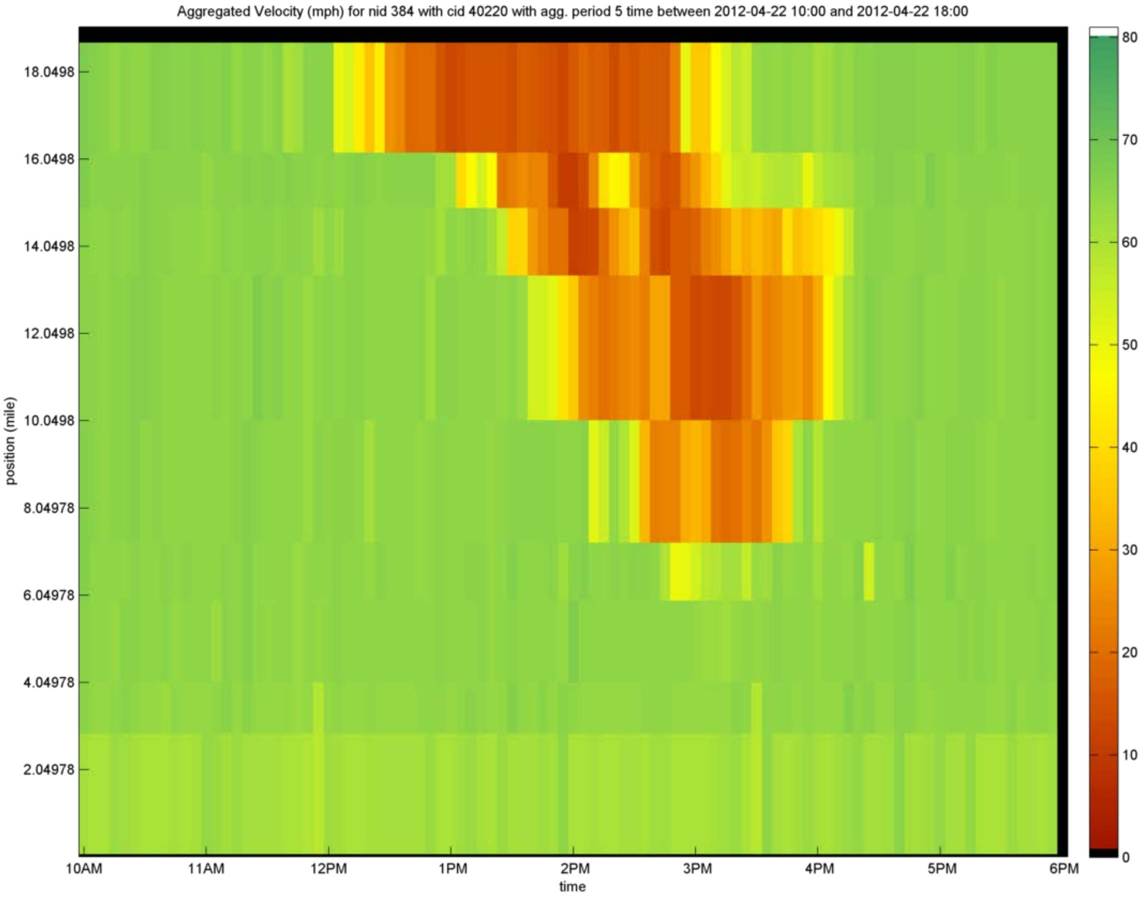


Figure 5-75: Average velocities on April 22 calculated from the reconstruction of vehicle trajectories using probe data. The velocities are calculated over 5-minute intervals along each of nine routes between successive Bluetooth detectors.

9.4 METRIC VS. NUMBER OF PROBE UNIQUE POINTS

When measuring the performance of probe data (type A and B combined), a clear improvement is evident as the average number of unique probe devices increases. In Figure 5-76, for example, the MAPE drops from 0.37 (37% discrepancy from Bluetooth measurements) with no data to less than 0.15 (15% difference from Bluetooth) with 1.3 unique probe devices per five minutes per mile, and reaches the "noise floor" of measured travel time variability with 1.7 devices per five minutes per mile.

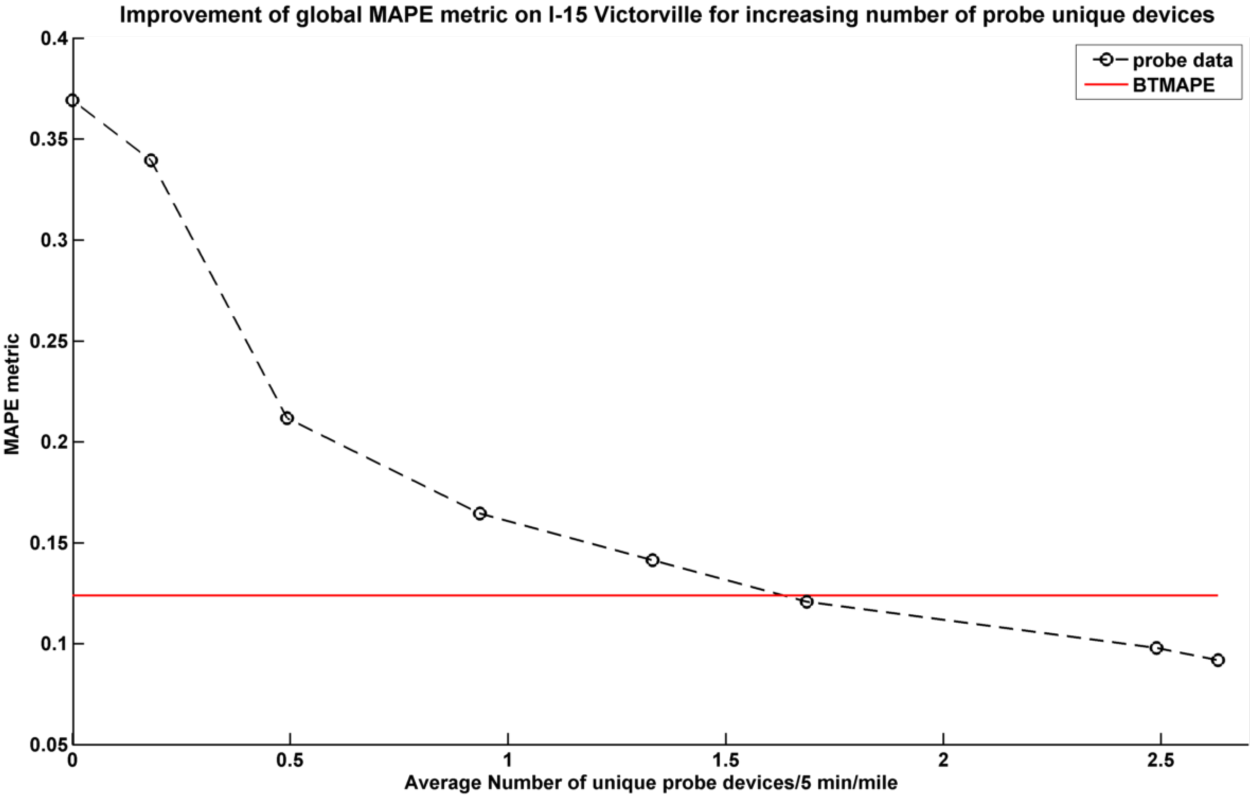


Figure 5-76: Trend of global MAPE vs. probe data (probe A and B)

Similar results can be seen for the PMATE metric in Figure 5-77 and the CCEC metric in Figure 5-78.

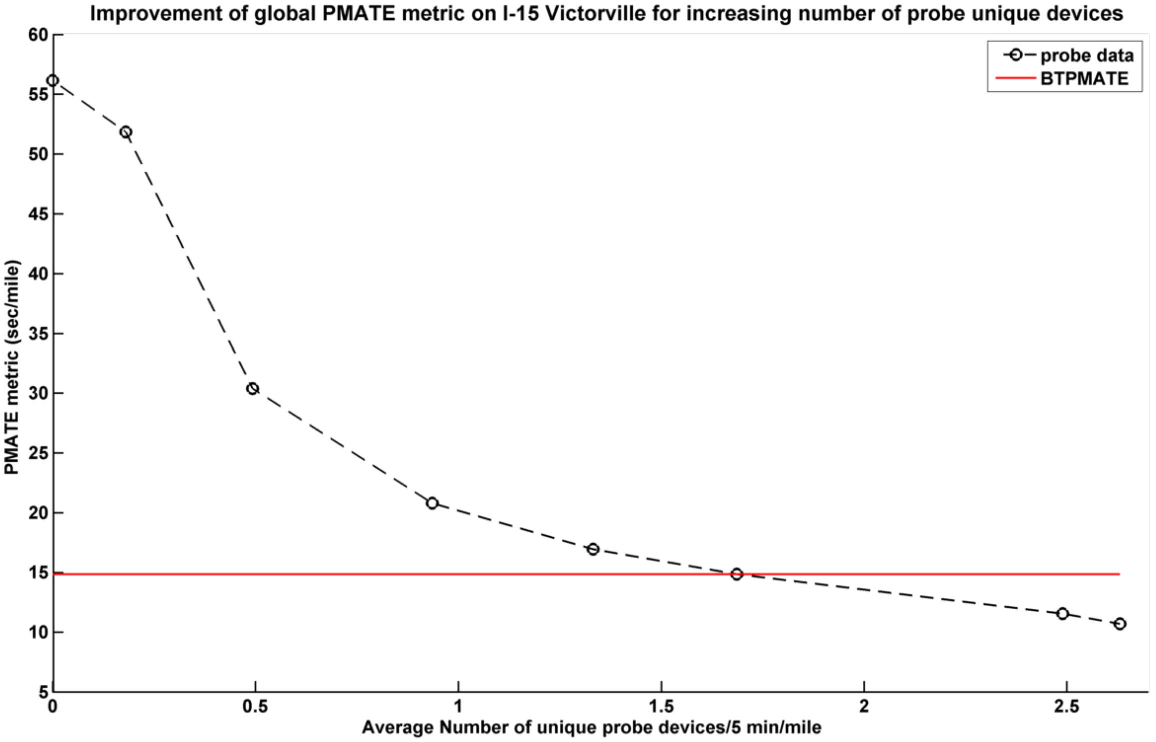


Figure 5-77: Trend of global PMATE vs. probe data (probe A and B)

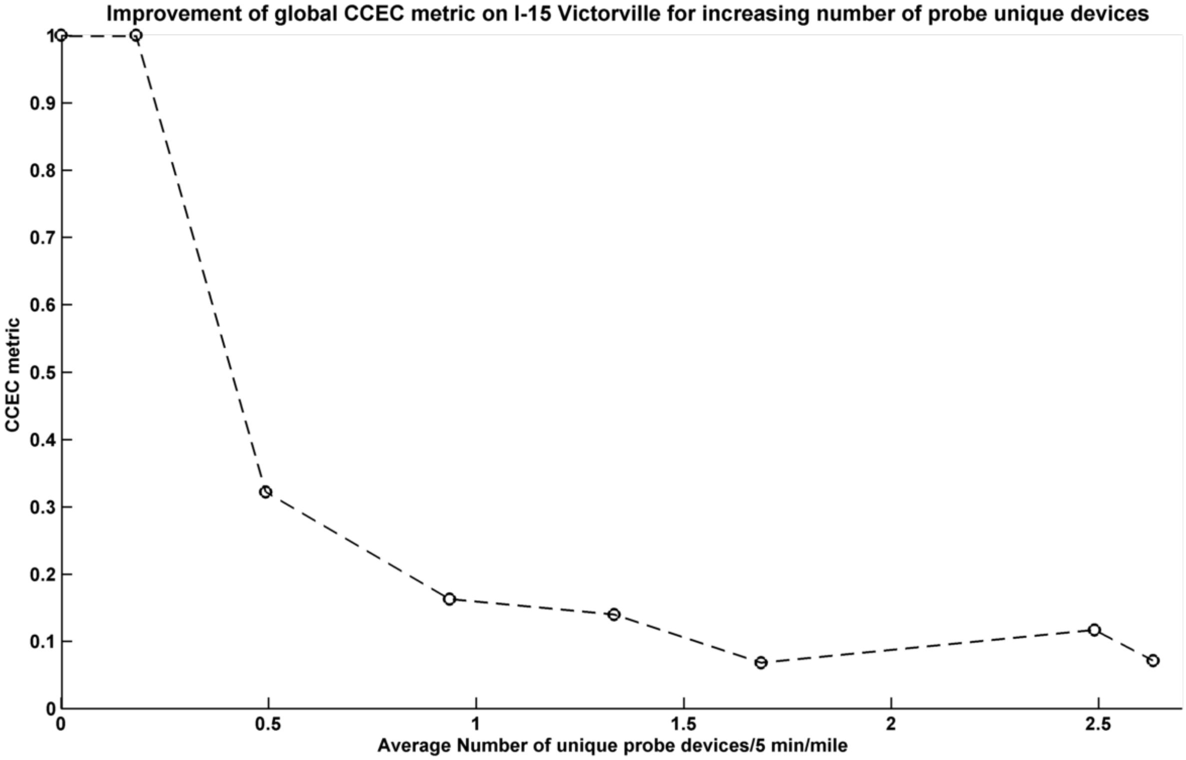


Figure 5-78: Trend of global CCEC vs. probe data (probe A and B)

10 SUMMARY AND FINDINGS

This section summarizes our data fusion investigation and findings:

- **Processing probe data.** We designed, implemented, and tested GPS probe data storage and data processing procedures. The processed data was assimilated in a flow model using a statistical filtering algorithm called the Ensemble Kalman Filter.
- **Assessing the data.** We defined pertinent data quality metrics and built a visualization tool that was used for descriptive and exploratory analyses of two months of probe data from two different providers. These analyses enabled us to compare the characteristics of the two different data sets and identify the main differences between them.
- **Increasing estimation accuracy.** Using Bluetooth travel time measurements as our estimate of ground truth, we were able to understand the resulting error (the discrepancy from Bluetooth measurements) as a function of the input data. Our analysis reveals that the accuracy of the model can be increased in two ways: (a) by adding more data (adding detectors or adding data points), and (b) using more than one type of data (fusing loop detector and probe data).
- **Adding data.** One way of adding probe data is to increase the frequency at which the GPS devices send their observations. A second way is to increase the number of unique devices that are detected, which in practice would mean increasing the penetration rate by flow of the number of vehicles (trucks, buses, taxis, etc.) transmitting GPS information. Our analysis shows that the number of unique devices is a better explanatory variable than the quantity of data points for explaining a performance difference between two data sets. This is an intuitive result because increasing the volume of data for the same vehicle does not add as much value to the model as increasing the number of vehicles. Too high frequencies can even be inefficient because of bandwidth and storage issues, so a reasonable frequency must be enforced, while trying to maximize the number of unique devices (which is a proxy of the penetration rate).
- **Using probe data alone.** We found that, for the sites we studied, the model achieved reasonable performance using probe data alone. Indeed, our results suggest that a similar level of accuracy can be obtained when using either probe data only (at 2012 penetration rates) or loop detector data only. This accuracy level also depends on the characteristics of the data set.
- **Loop spacing and position.** We performed a case study on the sensitivity of the performance metrics to loop detector spacing and position. Our results show that some loop detectors add more value to the model than others, and that installing fewer loop detectors but at locations pertinent from a traffic engineering point of view is more effective than installing more detectors just to enforce a minimal spacing between them. Although this case study relies on data from a limited time frame and area, we believe it is an interesting first step and that further research on detector spacing and position is needed to achieve more general results.

Chapter 6

Hybridization in Traffic Management Systems

Research into enhancing Traffic Management Systems to use multiple data sources was subcontracted as part of this task order. This chapter describes the subcontracting process and includes the subcontractor’s final report.

1 BACKGROUND: THE SUBCONTRACTING PROCESS

1.1 SOLICITING PROPOSALS

On February 28, 2011, a solicitation for proposals entitled “Data Fusion Research Subaward: Feasibility Study of Fusion Implementation in PeMS” was released by the California Center for Innovative Transportation (CCIT[6]). CCIT complied with the purchasing methodology outlined in University of California Business and Finance Bulletin BUS-43, particularly Part 3, §III, which refers to common goods, materials, and services over $100,000 in value. The solicitation document was made available to potential bidders by public release on CCIT’s website and by direct emails to several vendors identified as having previously engaged in related work.

CCIT indicated that it wished to “engage the services of an experienced consultant to provide research services to examine how purchased traffic data can be integrated into Caltrans' Performance Measurement System (PeMS) and then modify the PeMS Research system located at UC Berkeley to provide this functionality”.

1.2 VENDOR INQUIRIES

CCIT offered potential bidders the opportunity to ask questions about the project prior to submitting proposals. These questions were due by February 25, 2011. Two sets of questions were received. The anonymized questions and CCIT’s responses were posted on March 15, 2011 on CCIT’s website.

1.3 EVALUATING PROPOSALS

The due date for the proposals was March 21, 2011; the only proposal received by that date was from Berkeley Transportation Systems (BTS), with an anticipated total cost of $499,986.56 (USD).

On March 23, 2011, the proposal was reviewed by a team of CCIT staff, including principal investigator Alex Bayen and engineers Joe Butler, Ali Mortazavi, and Nazy Sobhi. They determined that the proposal substantially met the technical requirements and that the cost was acceptable. The BTS proposal was therefore determined to be the lowest bid meeting specifications, and, hence, BTS was awarded the work. (Vendor scoring criteria had been prepared on or about February 4, 2011, using a cost per quality point basis. The RFP, BTS proposal, scoring criteria, and contract are included in the Supporting Documents for Task Orders 1 and 2, provided to Caltrans in electronic format along with this report.)

The rest of this chapter presents BTS’ final report for the subcontract.

---

[6] CCIT has since merged with Partners for Advanced Transportation Technology (PATH) and operates under that name.

2 INTRODUCTION

2.1 HISTORY

The Partners for Advanced Transportation Technology (PATH) organization of UC Berkeley issued an RFP that described the need to integrate probe data into PeMS. Berkeley Transportation Systems (BTS), responded to the RFP, as it had expertise with the PeMS product. BTS was awarded the contract, and the four bullet points below briefly summarize important events and direction changes that occurred as part of the project.

1. Initial contract deliverables agreed on 3/18/2011.
2. BTS was acquired by Iteris Inc. in November 2011.
3. Given the passage of time between contract agreement and work start, PATH requested Iteris Inc. to contribute to some other tasks, with higher priorities that were not specified in the original RFP.
4. Caltrans agreed that PATH could direct Iteris Inc. as needed.

2.2 LEARNING FROM MOBILE MILLENNIUM

A great deal was learned by building Mobile Millennium. Iteris Inc. was asked to address the following requirements as part of the Data Fusion Research Subcontract (DFRS) deliverables:

1. All Highway modeling should be integrated into one project. Previously, Highway Forecast and Highway (Estimation) were separate.
2. Traffic engineers need to be able to run simulations without being Java programmers. Thus, the system should be customizable with respect to the additional functionality enumerated below.
3. Additional functionality is needed to support new research areas. In particular:
    a. The Cell Transmission Model (CTM) must support velocity, density, and augmented (velocity and density) data.
    b. Links must be capable of having unique fundamental diagrams.
    c. Links must support changing fundamental diagrams based on date/time.
    d. CTM must support the introduction of noise into ensemble elements.

3 OBJECTIVES

The main theme for the Data Fusion Research Subtract (DFRS) is to support the reliable processing of crowd-sourced GPS data and displaying of the results. The Grand Unification of Highway (a.k.a the Freeway project) contains the code that reads either probe data or roadway-embedded loop data to produce roadway speed and/or density estimates. The Data Quality Tool produces line charts and heat map representations of GPS probe data.

One possible use of the two tools together is to begin with the Data Quality Tool to gather an understanding of the crowd-sourced GPS data, followed by using the Freeway codebase to process them to produce roadway speed estimates. One possible learning is to understand how the quality of the probe data affects the roadway estimates.

3.1 GRAND UNIFICATION OF HIGHWAY

Iteris was asked to construct a research platform that unified the functionality implemented in Mobile Millennium Highway Forecast and Mobile Millennium Highway (Estimation). Investigators need to be able to configure the platform with various combinations of the parameters below:

1. Mode:
    a. Historical – Support repetitive running of a simulation where measurement data has been previously stored in a database.
    b. Live – Support running a model using live data.
2. Workflow:
    a. Forecast (CTM only)
        i. Database-resident initial condition
        ii. Create a default initial condition.
    b. Estimation (CTM and Ensemble Kalman Filter (EnKF))
3. CTM: Fundamental Diagram combinations:
    a. ρ-CTM: Greenshields
    b. ρ-CTM: Linear Hyperbolic
    c. ρ-CTM: Newell-Daganzo
    d. ρ-CTM: Additive Velocity Function Noise
    e. v-CTM: Greenshields
    f. v-CTM: Linear Hyperbolic
    g. Augmented (ρ+ v) CTM: Greenshields
    h. Augmented (ρ+ v) CTM: Linear Hyperbolic
    i. Augmented (ρ+ v) CTM: Newell-Daganzo
    j. Augmented (ρ+ v) CTM: Additive Velocity Function Noise

   k. Augmented (v+ρ) CTM: Greenshields
   l. Augmented (v+ρ) CTM: Linear Hyperbolic
   m. Augmented (v+ρ) CTM: Additive Velocity Function Noise

4. EnKF Types:

   a. Simple Average
   b. Global Jama
   c. Local Jama
   d. Global Blas
   e. Local Blas

5. Feed:

   a. Loops (PeMS Collected data)
   b. Probe A
   c. Probe B

3.1.1 DEVELOP PRODUCTION-QUALITY HIGHWAY ESTIMATION SYSTEM

High code quality in the area of numerical correctness is desired for the Freeway system, to gain confidence in the most important numerical components:

1. Cell Transmission Model (CTM)

2. Ensemble Kalman Filter (EnKF)

They were built into two technologies and their respective results compared. The two technologies are:

1. Java, which was chosen because it can be easily run on most computer systems in a Java Virtual Machine.

2. Mathematica, which was chosen because a player is available at no cost that allows running the implementation on most common computer platforms. The Expertise to complete this step was also available without training, which greatly reduced the effort of the implementation.

Comparison of the results of new system to Mobile Millennium for preexisting networks and time intervals could not be completed as hoped due to differences in the models.

3.2 EXTEND PROBE DATA QUALITY VISUALIZATION TOOL

PATH had a prototype Data Quality Tool that displayed crowd-sourced GPS data but which exhibited quality problems and functional deficiencies. The quality problems were addressed as they were observed, and the following functional enhancements were desired:

- Display PIF Data for Probes A and B
    - Develop aggregations by probe vendor to speed up the user experience. It often took many minutes for graphs to appear after the user had requested them. The goal was to have graphs to appear in less than 30 seconds.
    - Display location of probe measurements on a map.
- Additional Dimensions for Ad-Hoc Analysis
    - Time Coverage
    - Unique Devices
    - Sampling Rate
    - Date of Week Aggregates

4 GRAND UNIFICATION OF HIGHWAY A.K.A. THE FREEWAY PROJECT

The Grand Unification of Highway, hereafter referred to as Freeway, is an easily configured system that meets the objectives outlined in the first section. All classes of the Freeway Project can be found in the edu.berkeley.path.freeway package. A researcher can enter the following parameter values in one of two ways:

1. A JSON-formatted file with the filename passed into the JVM when Freeway starts.
2. Saved into a FreewayRunConfig table in the VIA database.

4.1 CONFIGURATION

All input parameters are passed in as objects found in the edu.berkeley.path.model_elements package. All classes in this package are serialized in an Oracle database, or a JSON file allowing saved parameter sets to be easily recalled as needed. The classes used directly for the configuration of Freeway are:

1. FreewayRunConfig

Units:

1. All time durations are in milliseconds.
2. All date-times are in milliseconds elapsed since midnight, January 1, 1970. For a further discussion of date-times, please examine the javadoc for java.util.calendar
3. All density means and density standard deviations are in units of vehicles/meter.
4. All velocity means and velocity standard deviations are in units of meters/second.

Below is a list of parameter names and value constraints. The parameter names can be identifiers in a JSON file or column names in the database-resident table. Each parameter has a suggested value that can be used by first-time users to maximize the probability that their configuration will run successfully.

FreewayRunConfig

id, name

These two parameters are not consumed by the Freeway code, but are present for people to use as labels useful to easily differentiate between different JSON files.

runMode

- "HISTORICAL" - Load all sets (FDs, Demands, Split Ratios) in advance, run five minutes per miles as fast as possible

- “LIVE” - Run at real-time speed, i.e., each 6-second CTM step takes 6 seconds.
    - Currently loads all sets (FDs, Demands, and Split Ratios) once in advance. Later, reload these periodically (to account for live calibration, etc.).
- “HISTORICAL_LIVE” - Use historical data, but work at real-time speed like live mode.
    - The advantage of using this mode over HISTORICAL is that the system load is reduced so that it can be run alongside a LIVE run.
    - The disadvantage of using this mode over HISTORICAL is that the run will require more elapsed time to complete. A run of one hour’s historical data will require one hour of elapsed time.
- Suggested Value: HISTORICAL – Please verify that data exists for the time interval configured below.

workflow

- “ESTIMATION” - Run CTM+EnKF
- “FORECAST” - Run CTM only
- “ESTIMATION_FORECAST” - CTM+EnKF and periodically spin off CTM-only forecast runs based on the latest estimated state
- Suggested Value: ESTIMATION

ctmType

- “Density” - rho-CTM
- “Velocity” - v-CTM
- “DensityVelocityFusion” - rho-CTM with state vector augmented by velocity
- “VelocityDensityFusion” - v-CTM with state vector augmented by density
- Suggested Value: “Velocity”

ensembleSize

- Range: integer x >= 1
- Suggested Value: 100 – Lower values run in less elapsed time.

fdType – Functional form of fundamental diagram. This determines which/how the FD parameters are used on each link.

- “Greenshields” - 2 parameter model using jam density and free-flow speed
- “DaganzoNewell” - 3 parameter continuous Daganzo-Newell model using jam density, free-flow speed, congestion wave speed

- “Smulders” (closest to MM equivalent) - 4 parameter Smulders (sometimes called “Linear hyperbolic”) model using jam density, free-flow speed, critical speed, congestion wave speed
- Suggested Value: “Smulders”

additiveModelNoiseMean, additiveModelNoiseStdDev

- Mean & standard deviation for Gaussian noise that will be added to the state vector after each CTM step.
- Range: -infintiy < mean < infinity, stdDev >= 0

additiveVelocityFunctionNoiseMean, additiveVelocityFunctionNoiseStdDev

- For creating a “noisy” velocity function. Mean & standard deviation for Gaussian noise that will be applied to the velocity function of each link during each CTM step.
- Not applicable for v-CTMs since a “noisy” velocity function defined this way is not an invertible velocity function.
- Range: -infinity < mean < infinity, stdDev >= 0

dtCTM

- CTM five minutes per mile in milliseconds. The smallest five minutes per mile in the system.
- Range: x > 0
- Suggested Value: 6000

dtOutput

- Time interval between generating and writing reports.
- Should be an integer multiple of dtCTM.
- Range: x > 0, x >= dtCTM
- Suggested Value: 30000 (5 minutes)

timeBegin, timeEnd

- Time range for “HISTORICAL” and “HISTORICAL_LIVE” mode. Sets (“profiles”) (demand, FD, split ratio) are preloaded to cover this time interval.
- 0 < timeBegin < timeEnd < infinity
- Suggested Value: none – Verify that data exists for the time interval specified.

enkfConfig

- All parameters specific to EnKF
- Can be null if using “FORECAST” mode

- **enkfType**

- MM-equivalent selection of EnKF implementations.
    - "LOCALJAMA", “LOCALBLAS” - EnKF with localization approximation using Jama or Blas library
    - “GLOBALJAMA”, “GLOBALBLAS” - Standard global EnKF using Jama or Blas library
    - “SIMPLEAVERAGE” - Non-EnKF (estimate is just average of measurements)
- Note: The EnKF localization approximation is not implemented yet, so the local solvers fall back to the global solvers.
- Suggested Value: “GLOBALJAMA”

- dtEnKF

- Time interval between EnKF runs. Should be an integer multiple of dtCTM.
- Range: x > 0, x > dtCTM
- Suggested Value: 30000

- includePeMS true/false**, pemsNoiseMean, pemsNoiseStdev**

- Whether to include PeMS measurements and what noise properties to assume for them.
- Range: -infinity < mean < infinity, stdDev >= 0
- Suggested Value: true, mean=0, stdDev=4.0 -- Only select ‘true’ after verifying that PeMS data is available for the roadway and time interval being simulated.

- PemsBlackList

- List of integer VDS IDs to exclude even if they appear in the scenario’s SensorSet. This parameter exists to give the experimenter the ability to vary the number of sensors used in simulations.
- Suggested Value: empty list

- probeProbabilityThreshold

- Accept map-matched probe measurements only if they have at least this probability assigned by PIF.

- Range: 0 <= x <= 1
- Suggested Value: 0.7

- probeSpeedThreshold

- Accept probe measurements only if they are above this speed. To filter out parked cars, etc.
- Range: x >= 0
- Suggested Value: 1.0

- pifRunId

- Load map-matched probe measurements only with this run ID that corresponds to a particular run of PIF.
- Suggested Value: Interrogate the database to learn which values are valid for the time interval and roadway used in the simulation.

- includeNavteq (true/false)**, navteqPercentage, navteqNoiseMean, navteqNoiseStdev**

- Not yet supported
- Whether to include Navteq probe measurements, what fraction of them to include, and what noise properties to assume for them.
- Range: 0 <= percentage <= 1, -infinity < mean < infinity, stddev >= 0
- Suggested Value: false, mean=0, stdDev = 4.0 -- Only select ‘true’ after verifying that data is available for the roadway and time interval being simulated.

- includeTelenav (true/false)**, telenavPercentage, telenavNoiseMean, telenavNoiseStdev**

- Not yet supported, but the names reserved for use in setting up future configurations.
- Whether to include Telenav probe measurements, what fraction of them to include, and what noise properties to assume for them.
- Range: 0 <= percentage <= 1, -infinity < mean < infinity, stddev >= 0
- Suggested Value: false, mean=0, stdDev = 4.0 Only select ‘true’ after verifying that data is available for the roadway and time interval being simulated.

- useLocalization (true/false)**, localizationDistance**

- Not yet supported, but the names reserved for use in setting up future configurations.
- Whether to use localized approximation to EnKF

  - Range: 0 < distance < infinity
  - Suggested Value: false, distance = 100.0

estimationForecastConfig

- Parameters specific to estimation-forecast runs in which period forecast runs are spawned off the latest estimated state
- Can be null

- dtEstimationForecastSpinoff

  - How often to spin off forecast runs from the latest estimated state.
  - Should be an integer multiple of dtEnKF. E.g. every minute.
  - Range: x > 0, x >= dtCTM, x >= dtEnKF
  - Suggested Value: 300000 (5 minutes)

- forecastDuration

  - How far ahead to forecast in each spun-off forecast run.
  - Range: x > 0
  - Suggested Value: 1800000 (30 minutes)

- dtEstimationForecastReport

  - How often to record the forecast statistics.
  - Range: x > 0, x >= dtCTM
  - Suggested Value: 300000 (5 minutes)

initialState, initialEnsembleState

- FreewayCTMState or FreewayCTMEnsembleState object giving the initial state or ensemble of states.
- See example freeway/examples/HistoricalForecastLocalWithInitialState.json. Can be found in the CALPATH/FREEWAY repo.
- Can be null, in which case initial state is determined using initialDensityFraction below.
- Suggested Value: null, null

initialDensityFraction

- Initial average density on links as a fraction of jam density. Used to randomly initialize the ensemble of states in the absence of any explicit initial state.
- Range: 0 <= x <= 1
- Suggested Value: 0.01

initialStateUncertainty

- Standard deviation of Gaussian noise (with mean = 0) that should be added to the initial ensemble of states.
- For rho-CTM, units are vehicles per meter. For v-CTM, units are meters per second.
- Range: x >= 0
- Suggested Value: 0.0

reportStatisticsAfterCTM, reportStatisticsAfterEnKF

- Which stage of each five minutes per mile to generate a statistical report.
- In Freeway at present, at most one should be = true.
- Suggested Value: CTM = false, EnKF = true

reportEnsembleAfterCTM, reportEnsembleAfterEnKF

- Which state of each five minutes per mile to report the full ensemble state.
- In Freeway at present, at most one should be = true.
- Suggested Value: Both = false

reportStatisticsHistory

- Whether to write a final report to JSON file containing snapshots of statistics report throughout entire run.
- This is automatically set to true by forecast runs spun off from estimation-forecast runs.
- Suggested Value: false

reportToDB

- Whether to write reports to their corresponding database tables.
    - For statistics report: LINK_DATA_TOTAL
    - For ensemble report: LINK_DATA_TOTAL_DEBUG

- Suggested Value: true

reportToDirectory, reportDirectory

- true/false, and string. Whether to write reports to a local directory, and which directory to use.
- Relative to startup directory.
- A subdirectory is created with name = the run ID assigned by the DB. For runs in which reportToDB is false, a run ID of 0 is used (can lead to files being overwritten).
- Suggested Value: false, “reports”

Parameter values used by Mobile Millennium

The following configuration settings are the values implemented in Mobile Millennium. They are listed here for use by anyone trying to approximate MM behavior:

```
ModelNoiseMean = 0.5
ModelNoiseStdev = 2.0
PemsNoiseMean = 0.0
PemsNoiseStdev = 4.0
NavteqNoiseMean = 0.0
NavteqNoiseStdev = 4.0
TelenavNoiseMean = 0.0
TelenavNoiseStdev = 4.0
PemsBlackList = null – All stations are operating correctly
TelenavPercentage = 100.0
NavteqPercentage = 100.0
UseLocalization = true
LocalizationDistance = 100.0
ConfidenceNoMeasurement = 0.0
ConfidenceDefault = 0.025
ConfidenceHasMeasurement = 0.5
ConfidenceMeasurementLifetime = 180.0
```

4.2 TESTING

A reliable Freeway project requires significant testing to guarantee robustness. Tests were developed to guarantee numerical correctness.

Numerical Correctness. The numerical correctness of the CTMs and the EnKF code was verified by comparing the Freeway outputs to two other systems:

1. Mathematica implementation of the CTM
2. Mathematica implementation of the non-localized EnKF

A version of the CTM and the EnKF were implemented in Mathematica to generate test vectors that were compared to the output of the Freeway code. The test cases developed:

1. Simple Junction – used to verify basic operation of the CTM
2. Non-localized EnKF – used to verify correct operation of the EnKF

5 PROBE DATA QUALITY VISUALIZATION

The Data Quality Tool displays graphs and maps of probe data feeds for a variety of metrics. The tool was enhanced in the following ways.

Correct Code Deficiencies

- Enabling/disabling controls based on available data
    - The initial version of the tool had a user interface issue where the configuration controls (e.g., Speed, Total transmission delay, etc.) were disabled because the UI code, erroneously, determined that data was not available for display. This defect was corrected.
- Map display of probe points
    - The initial version of the tool did not display data on the map present below the line chart. This deficiency was the result of missing configuration information about each of the study sites. Metadata was added, and the map behaved correctly.

Correct Performance Deficiencies

- Aggregated PIF Data by Study Area and Time
    - The initial version of the tool pulled raw data from the database and performed basic aggregation as part of the UI code. The in-line processing of the data resulted in multi-minute delays in report generation requests. To rectify this behavior, code was written to pre-aggregate the raw values. All report-requests are now returning within time intervals which are acceptable to report users.

Functional Extensions

- The following metrics are now supported:
    - Speed
    - Total transmission delay
    - Provider transmission delay
    - Space-time coverage
    - Time coverage – for raw data and map-matched data
    - Sampling rate – for raw data and map-matched data
    - Unique devices – for raw data and map-matched data

- Support Day-of-Week Dimension
    - Previously, all data within the requested time interval was displayed. The day-of-week dimension was added to allow users to analyze incoming data by the day of the week it was collected.

A user guide for the Data Quality Tool is included in the final report for Task Order 2, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods.*

6 PEMS RESEARCH INSTALLATION

An installation of PeMS was set up for future research use by PATH personnel. Note that the configuration is abbreviated to minimize the hardware needed. PATH supplied a two-core system with 16 GB of RAM and 1 TB of disk for this purpose.

Historical data from Caltrans Districts Four and Eight covering the time interval of January 2012 to June 2012 was loaded into the PeMS database to cover the roadway segments that were extensively analyzed during Task Orders 1 and 2. The data was processed to produce the aggregations needed to drive PeMS reports.

A live feed covering District Four was also set up. The live data is sent from the Caltrans PeMS instance located in the Sacramento area, which has been configured with an additional Value-Added Reseller (VAR) feed, to the PATH research instance.

7 CONTRACTUAL WORK PACKAGES

Chapter 5 of the Data Fusion Research Subcontract solicitation describes five Data Fusion, i.e., Documentation, “work packages” and three Software “work packages”. While the work completed by Iteris was redirected as described in section 2.1, the text below describes how the work performed applies to the proposed work breakdown structure.

7.1 DATA COLLECTION RESEARCH

A major aspect of this step is mapping of the raw probe data to a section of a freeway. As PATH has already built the Path Inference Filter (PIF) that supplies this functionality, Iteris Inc. was encouraged to use the existing code.

7.2 DATA CLEANING RESEARCH

The PIF, mentioned in the previous section, natively provides the Data Cleaning function. Again, Iteris Inc. was encouraged to use the existing code. Note that Iteris Inc. contributed to this activity by including configuration flexibility in the Freeway project to support further refinement of Data Cleaning, by allowing researchers to request varying percentages of probe measurements to be utilized by the models. Thus, the optimal probe measurement density can be determined for each feed type.

7.3 DATA FUSION RESEARCH

The Freeway project is capable of being configured to take in multiple data sources and return a single velocity and/or density measurement for a roadway. Please see section 4.1, noting the values selected for the following parameters:

- includePeMS
- includeNavteq
- includeTelenav

7.4 DATA USE RESEARCH

Iteris Inc. extended a Data Quality Visualization prototype that was developed as part of the Mobile Millennium project. Section 5*, Probe data quality visualization*, describes the work performed.

7.5 SUSTAINABILITY RESEARCH

The proposed sustainability areas included Scalability, Data Flow Management, and Historical Data Storage. These were not addressed, as Code Quality required attention first. Thus, this study area was deferred.

7.6 PROTOTYPE SOFTWARE DESIGN DOCUMENT

The Freeway Project produces velocity estimates that could be consumed by PeMS and displayed in PeMS graphs along with loop velocity estimates. The structure and fundamental operation of the Freeway Project was documented and delivered to PATH.

7.7 PROTOTYPE SOFTWARE RESEARCH

The Freeway Project contains the Highway Estimation algorithm. It was written, tested, documented, and delivered to PATH.

7.8 PROTOTYPE SOFTWARE DATABASE SCHEMAS

Tables needed for all phases of crowd-sourced GPS probe processing were designed and implemented in the VIA database. The tables contain:

1. Raw Probe Measurements
2. Filtered and cleaned data as a result of the PIF process. These are the input tables for the PIF Freeway Project.
3. Freeway output tables containing the velocity and density estimates for the roadways of interest.
4. All tables required for the successful operation of a PeMS instance. They are currently being used by the PATH research instance of PeMS.

8 PROBE DATA IN PEMS – ESTIMATED DEPLOYMENT EFFORT

The work completed during this contract has removed significant technical risk from the objective of using probe data in PeMS. The tasks below, and work estimates for each, describe the remaining work needed for deployment across the state of California.

8.1 ASSUMPTIONS

1. The UCB PATH-developed Path Inference Filter (PIF) is open source and available for royalty-free use in the system.
2. The Cost of the probe data is not included in the estimate.
3. The cost of the map is not included in the estimate. An open source selection, for example Open Street Maps (OSM), would force this cost to zero.
4. The cost of the Oracle 11g R2 database license is not included in the estimate.
5. Real-Time is defined to indicate that data collected every 30 seconds can be processed within the subsequent 30 seconds of elapsed time.

8.2 ASSUMED HARDWARE REQUIREMENTS

1. System(s) to process incoming probe data
    a. Run PIF, associating probe measurements with their corresponding freeway links
2. System(s) to run Estimation and Forecast
3. System(s) for initial, and ongoing, CA Highway system Calibration
4. System(s) to run PeMS – Is this installation parallel to the existing deployment?

8.3 DEVELOPMENT TASKS

1. Design Deployment Architecture
    a. Deliverable: Document describing the Interfaces between processing elements.
    b. Effort estimate: 1 staff month

2. Determine the hardware requirements for each of the systems described in the architecture above.
    a. Deliverable: Document describing:
        i. The hardware requirements for a system supporting the state of CA.
        ii. Model description and measurements that justify the system requirements
    b. Effort estimate: 1 staff month

3. Design and Deploy framework
    a. Deliverable: Running system consisting of shell classes that implement the interfaces described in step 1 above. 'Dummy' traffic flows through the entire stack
    b. Effort estimate: 2 staff months

4. PIF – One district
    a. Deliverables:
        i. The road network in the covered district has been codified so PIF can run. The form of the codification is described in the architecture produced in step 1.
        ii. The parallelized PIF process running for one district. Transform of incoming probe data to map-based links as described in the architecture document produced in story 1.
    b. Effort estimate: 3 staff months

5. Freeway – Estimation – one district. Note that Forecast is not included in this story.
    a. Deliverables:
        i. A real-time Estimation processing element that supports fusion of GPS Probes and Loop Detectors, perhaps, including:
            a. Parallelizable CTM (possibly by ensemble element)
            b. Parallelizable EnKF (possibly by ensemble element)
            c. Parallelizable Measurement Loader capable of supporting probe data and in-road loop detectors
    b. Effort estimates:
        i. Parallelizable CTM: 6 staff months
        ii. Parallelizable EnKF: 6 staff months
        iii. Parallelizable Measurement Loader: 3 staff months

6. PeMS Integration – one district
    a. Deliverable:
        i. PeMS reports containing Freeway Estimation results. Enumerated list of reports is needed for more precise estimate.

   b. Effort estimate: 6 staff months

7. Scalable PIF – All of CA

   a. Deliverable: Replicate PIF environments as required to cover all roadways under Caltrans jurisdiction.
   b. Effort estimate: 2 staff months per district; 14 staff months total

8. Scalable Freeway Estimation – All of CA

   a. Deliverable: Replicate Freeway Estimation environments as required to cover the roadways under Caltrans jurisdiction.
   b. Effort estimate: 1 staff month per district; 7 staff months total

9. PeMS Integration – All of CA

   a. Deliverable: PeMS database can successfully receive data from all running Freeway Estimation modules concurrently.
   b. Effort estimate: 2 staff months

10. Calibration Basic Science and System Development

    a. Deliverables:
       i. Basic science to determine algorithms to estimate calibration parameters
       ii. A system supporting operator interaction to deliver calibration estimates
    b. Effort Estimates:
       i. Basic science: 12 staff months
       ii. Calibration system: 24 staff months

11. Calibrate CA Highway – one district

    a. Deliverable: Estimates for Calibration parameters for the area covered:
       i. Split ratios for all junctions as a function of time/day
       ii. Ramp demand as a function of time/day
       iii. Fundamental Diagram selection as a function of speed limit (FF speed)
    b. Effort estimates:
       i. Calibration parameters for the district of interest: 9 staff months

12. Calibrate CA Highway – All of CA

    a. Deliverable:
       i. Estimates for Calibration parameters for all the remaining seven districts.
    b. Effort estimate: 7 districts * 6 staff months each = 42 staff months

8.4 MAINTENANCE (RECURRING) TASKS

Once the system is deployed, the following tasks will be completed periodically throughout the deployment lifetime.

1. System Maintenance
    a. Effort estimate: 12 staff months per year
2. Calibration Maintenance:
    a. Deliverable: Updated Estimates for all Caltrans Districts on an ongoing basis.
    b. Effort estimate: 1 staff month per district per revision interval.

8.5 DEPLOYMENT TASK ESTIMATE SUMMARY

Deployment Task	**Effort Estimate (staff months)**
Design Deployment Architecture	1
Determine Hardware Requirements	1
Design and Deploy Framework	2
PIF – one district	3
Freeway Estimation – one district	15
PeMS Integration – one district	6
Scalable PIF – All CA	14
Scalable Freeway Estimation	7
Scalable PeMS – All CA	2
Calibration – Basic Science and System Development	36
Calibrate CA Highway – one district	9
Calibration of all CA Highways	42
Total	138 = 11.5 staff years

Chapter 7

Conclusion

1 SUMMARY OF THE PROJECT

Faced with the rising cost of maintaining its traditional traffic data collection systems, such as loop detectors installed at fixed locations, and recognizing the growing prevalence of commercial traffic data sources, Caltrans has begun looking into procuring good-quality data from third-party sources and integrating that data into Caltrans' existing systems. As part of that effort, PATH undertook this pilot procurement project to explore the feasibility of purchasing third-party probe data (unaggregated GPS point speeds) from the commercial sector and fusing it with Caltrans' existing data for the purpose of estimating travel times. The intent was to demonstrate an efficient and cost-effective use of alternative traffic data sources to complement the detection systems currently installed and operated by Caltrans.

Because the aim of the project was both to acquire probe data and to use it to estimate travel times along the highway, the investigation pursued a range of objectives in both engineering and business, including:

- **Assess current practices** in the traffic management community by interviewing stakeholders from public agencies and industry and reviewing previous procurement efforts from both Caltrans and other state DOTs. What we found was a market undergoing rapid evolution and a widespread interest among stakeholders in integrating probe data with loop detector data. Stakeholders were seeking both accurate speed and travel time information (in addition to the volume and occupancy data currently available from loop detectors) and the ability to characterize corridor performance in more precise and comprehensive ways.

- **Develop a process for obtaining probe data** from the commercial sector. We issued an initial Request for Information (RFI) to gauge vendors' interest, expertise, and experience in supplying data. Based on those responses, as well as what we had learned from other transportation stakeholders, we went through the process of defining the data to be acquired (unaggregated GPS point speeds) and soliciting proposals from potential suppliers through a Request for Proposals (RFP). The RFP required a balance of scientific rigor and simplicity—specific enough to get the data we wanted but not so complex that it discouraged vendors from responding. We then established evaluation criteria, reviewed the responses, selected vendors, and contracted for product delivery.

- **Identify the highway segments** for which data would be procured. We were seeking locations where traffic was both variable and representative so that the traffic state was not obvious, all ranges of speeds could be assessed, and the observations could be considered applicable to places with similar traffic phenomena. Using historical data from Caltrans, as well as historical and live data from Google Traffic, we found three segments with suitable traffic conditions: I-880 in the Bay Area, I-15 in Ontario, and I-15 in Victorville. To get independently observed travel times to use as a benchmark, we installed Bluetooth sensors at the selected sites. This gave us a reasonable estimate of the actual traffic conditions (the "ground truth") that the purchased data could be compared to.

- **Integrate the data** in a flow model to estimate travel times. The probe data from each vendor was passed through a series of filtering steps which verified that the required data fields were present, removed faulty data, and mapped the point-speed locations to the road network. The data was then fed into the Mobile Millennium highway model which generated velocity maps and travel times. These could then be compared to the travel times collected by the Bluetooth sensors to see how closely the probe data matched the estimated ground truth. We also combined the purchased data with data from Caltrans’ loop stations to see the extent to which probe data could supplement or supplant loop detector data.

2 SUMMARY OF FINDINGS

Data procurement

While the market is continually evolving, unaggregated probe data is available and can be procured successfully through a Request for Proposals (RFP) process, as outlined in Chapter 4. Although limited in scope and aimed at specific targets, the RFP we issued stimulated the market to respond and revealed that vendors had both the desire and capability to supply the requested data. Our initial Request for Information (RFI) elicited interest from ten vendors, four of whom subsequently submitted proposals in response to our RFP. Of those four, two were selected to supply data for the project.

Speed data and travel times

Commercially available GPS point-speed data is usable for the intended application (estimating speed data and travel times over a highway network) and can be successfully processed with the Mobile Millennium system to map velocities and compute travel times. The Path Inference Filter (described in the Task Order 2 final report, *Hybrid Traffic Data Collection Roadmap: Objectives and Methods*) can effectively filter both high-frequency and low-frequency GPS data to remove duplicates and outliers and can accurately map the vehicles and their trajectories to the road network.

Data quality

The quality of purchased probe data can be measured and compared to ground truth. The metrics described in Chapter 5 quantify the difference between probe data and the Bluetooth measurements used for ground truth, while the Data Quality Tool described in the final report for Task Order 2 can be used to examine the probe data characteristics directly. We also found that when using probe data with a flow model on a highway, better performance was achieved by having less frequent data from a larger number of unique vehicles (a high penetration rate) than by having more frequent data from a smaller number of unique vehicles (a high sample rate). High-frequency data might be more useful for other applications, such as arterial estimation.

Data fusion

Probe data can be successfully fused with loop detector data, and meaningful comparisons can be assessed. The Mobile Millennium system can accept data from traditional sources (such as occupancy and counts from loop detectors) and point-speed measurements from providers of probe data. This enabled us to evaluate the performance of the data sources both individually and in combination. By reducing the amount of data from any particular source that we use with the model, we were able to test different percentages of each and evaluate the results with various proportions of probe and loop detector data.

Traffic estimation

The accuracy of traffic estimation can be improved by combining probe data with loop detector data; even sparse probe data is useful. When probe data and loop data are fused together, average travel times can be estimated at an accuracy within the bounds of driver variability. For the purpose of estimating travel times, probe data can be used as a substitute for data from loop detectors.

Confidence in the model

The quantitative and graphical results of the research give us confidence in both the modeling approach to roadway estimation and the effectiveness of the Mobile Millennium highway model for assessing and fusing procured data.

3 IMPLICATIONS AND FUTURE DIRECTIONS

Procuring commercially available probe data is just one of many important steps toward improving mobility in California by leveraging new data sources. This pilot project has shown that third-party probe data can be purchased commercially, fused with Caltrans' loop detector data, and used to improve traffic estimation along California's roadways without building new detector infrastructure. Probe data could even potentially fill in gaps where no loop detectors exist and speed data alone is sufficient.

Several implications follow from these findings:

Reduced dependency on loop data

While current system control strategies, such as ramp metering, require density data, it seems difficult to significantly increase and maintain the quantity of loop detectors on California roads. At the same time, the penetration rate of probe data is continually increasing and far from reaching its limits. This represents a sea change in the types of data available for traffic management and offers the prospect of migrating away from exclusive dependency on loop detectors over time.

Outsourced data collection

Purchasing probe data from the commercial sector means, in effect, outsourcing the collection of traffic data. Any such undertaking comes with new risks (e.g., data quality, privacy protection, business continuity) which would need to be managed through, for example, a careful vetting and data acquisition process and data assessment tools, processes, and standards.

Redesigned information systems

The research work in this task order was predicated on having a data assimilation and state estimation system in place that would allow the implementation, testing, and analysis of data hybridization. This required:

- Building and calibrating a model to estimate speed and travel times from probe data
- Developing a set of methods to fuse probe and loop detector data
- Creating tools to visualize the data
- Testing the tools and methods on real data from pilot sites
- Building tools to determine the quality of the methods, models, and data

Creating this mathematical and technology infrastructure points the way to the redesign of information systems that would make it possible to implement data fusion and take full advantage of hybrid data in traffic management systems.

New detector strategy

While further research on the position and spacing of loop detectors is needed, our initial results suggest that using the most critical detectors (those that add the most information value) rather than enforcing a minimal spacing between detectors could help optimize the existing stock of detectors and allow Caltrans to selectively focus maintenance efforts or supplement loop data with probe data when certain loops fail.

Augmented traffic measurements

This project studied the use of probe data for estimating travel times. However, the enhanced modeling and estimation accuracy demonstrated by data fusion also lays the foundation for better control strategies and operational decision-making. Augmenting traffic volume measurement with probe data, for example, could be a fruitful area for research in the future. Fused loop and probe data (“hybrid” data) could thus provide a pathway to the development and use of additional traffic measurements, such as arterial estimation, origin-destination information, demand modeling, and, ultimately, integrated corridor management.

Broader potential

Being able to reliably purchase and use accurate traffic data could potentially enhance many areas of interest to Caltrans beyond the efficient flow of goods and people across California, including transportation safety, work zone safety, emergency services, and evacuation management. The successful procurement and assimilation of third-party data represents an important step toward those possibilities.

Bibliography

REFERENCES

[1] B. Greenshields, "A study of traffic capacity," in *Proceedings of the 14th annual meeting of the Highway Research Board*, 1934.

[2] S. Godunov, A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics, 1959.

[3] I. S. Strub and A. M. Bayen, "Weak formulation of boundary conditions for scalar conservation laws: an application to highway traffic modeling," *Int. J. Robust Nonlinear Control,* 2006.

[4] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, Springer, 2007.

[5] D. B. Work, S. Blandin, O. Tossavainen, B. Piccoli and A. M. Bayen, "A traffic model for velocity data assimilation," *Applied Mathematics Research eXpress,* 2010.

[6] G. Burgers, P. Jan van Leeuwen and G. Evensen, "Analysis scheme in the ensemble kalman analysis scheme in the ensemble kalman filter," *Monthly Weather Review,* vol. 126, no. 6, p. 1719–24, 1998.

[7] P.-E. Mazaré, O.-P. Tossovainen, A. Bayen and D. Work, "Trade-offs between inductive loops and gps probe vehicles for travel time estimation: A mobile century case study," 2008.

Appendix:

A Study of Traffic Variability for Building a Reference State

1 INTRODUCTION

This chapter describes a methodology used to investigate traffic variability, in a study conducted in the San Francisco Bay Area. The purpose of the study was to find relevant space and time windows to build a reference state for validating third-party data. The results informed our site selection for the probe data procurement pilot, both in the Bay Area and southern California.

The reference state, we believe, should be established where traffic is both variable and representative for the following reasons:

- We want to be able to assess the latency of the data feed. Consequently, we do not want to apply the validation process at a place where traffic is quite constant.
- We want a location where traffic is difficult to predict, so the periodicity of traffic cannot be exploited to provide a feed based on historical data.
- Such a place has to be representative of traffic to be able to extend the validation to places with similar traffic evolutions.
- The validation process has to be done for each range of possible speeds. We want to record data in all ranges of speed so places where congestion occurs can be exploited.

2 METRICS DEFINITION FOR TRAFFIC VARIABILITY

In this study, we had speed measurements from 114 radar detectors. The radar processed feed was used to analyze the variability of traffic in the Bay Area. In order to do so, three characteristics were defined and studied for the month of September 2010.

2.1 VARIABILITY DURING A DAY: σ

The speed was measured every day from 6 am to 12am (18 hours).

$V_{day\_d} = (v_{1d}, v_{2d}, \dots, v_{Nd})^T$ where N is the number of speed measurements we have at this location for the day considered. Since the radar detectors provide new points of data every minute, N is usually close to $18 * 60 = 1080$. So we have 1080 points per radar per day.

We computed the standard deviation of this speed every day.

$$\sigma_{day\_d} = \sqrt{\sum_{i=1}^{N} \left(v_{id} - \overline{V_{day\_d}}\right)^2}$$

$$\sigma_{day\_d} = \sqrt{VAR(V_{day\_d})}$$

σ is the mean of the standard deviations computed for all the n days of the month.

$$\sigma = \frac{1}{n} \sum_{d=1}^{n} \sigma_{day\_d}$$

Therefore, σ represents the variability of the speed during a day.

2.2 VARIABILITY BETWEEN DAYS

- $\bar{\boldsymbol{d}}$ -> mean of the L1 errors between days

The L1 errors between all possible combinations of two days d_i and d_j is computed. $\bar{d}$ is the mean of the L1 errors between days. The L1 error between two days indicates "how different from each other" the two days are as regard to the evolution of the speed during these days. As a consequence, $\bar{d}$ represents "how much" days can differ from each other at this particular location.

Mathematically we built the matrix of the e_{ij}:

First, a trapezoidal approximation is used to compute integrals of $V_{day\_i}(t)$ which is built from the vector with discrete values V_{day_d}.

$$e_{ij} = distance(day\ _i, day\ _j) = \left\|V_{day\_i} - V_{day\_j}\right\|_{L1} = \int_{t=6am}^{12am} \left|V_{day\_i}(t) - V_{day\_j}(t)\right|\ dt$$

$$E = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & \mathrm{e_{1n}} \\ \vdots & \ddots & \dots & e_{ij} & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & e_{ji} & \cdots & \ddots & \vdots \\ \mathrm{e_{n1}} & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

$$\bar{d} = \frac{1}{n(n-1)/2} \sum_{i=1}^{n} \sum_{\substack{j>i \\ j\le n}} e_{ij}$$

- $\boldsymbol{\sigma_{dist}}$-> standard deviation of the L1 errors between days

As explained above, the L1 error between two days indicates "how different from each other" the two days are as regard to the evolution of the speed during these days.

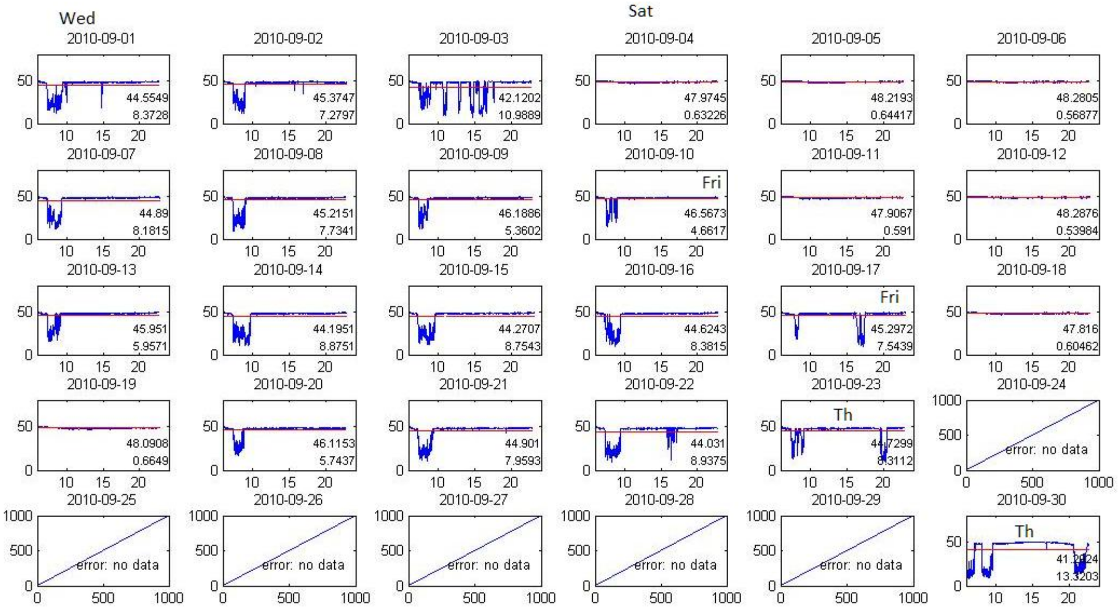
$$\boldsymbol{\sigma_{dist}} = \sqrt{\sum_{i=1}^{n} \sum_{\substack{j>i \\ j\le n}} (\ e_{ij} - \bar{d}\ )^2}$$

Consequently, $\boldsymbol{\sigma_{dist}}$ represents the variability between two days.

3 METHODOLOGY: FINDING CANDIDATES FOR THE LOCATION OF THE REFERENCE STATE

3.1 EFFECT OF WEEKENDS

A preliminary study showed that for a fixed location, the speed during the weekends is constant most of the time. Consequently, the weekends may increase the variability significantly between days as defined by the metrics above when traffic evolves during the weekdays. On the chart below, this fact can be observed.



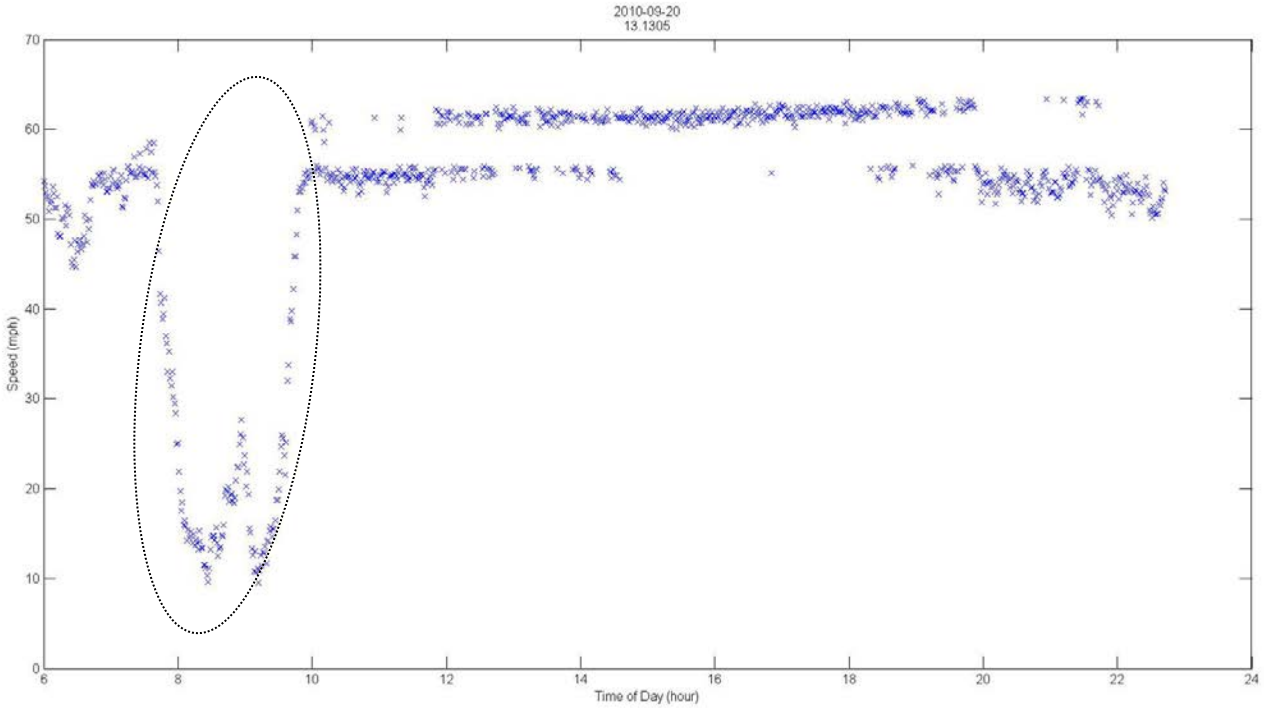
As we can see, the curve of the speed is flat for all Saturdays and Sundays. Consequently the L1 error for this location is going to be high though the speed of the other weekdays has similar evolutions.

As a conclusion, **the study of the variability of traffic between days should not take into account the weekends**. Our metrics were computed for all the days of the month without taking the weekends into account.

3.2 PENETRATION RATES OF TECHNOLOGIES AND IMPLICATIONS ON FLOW

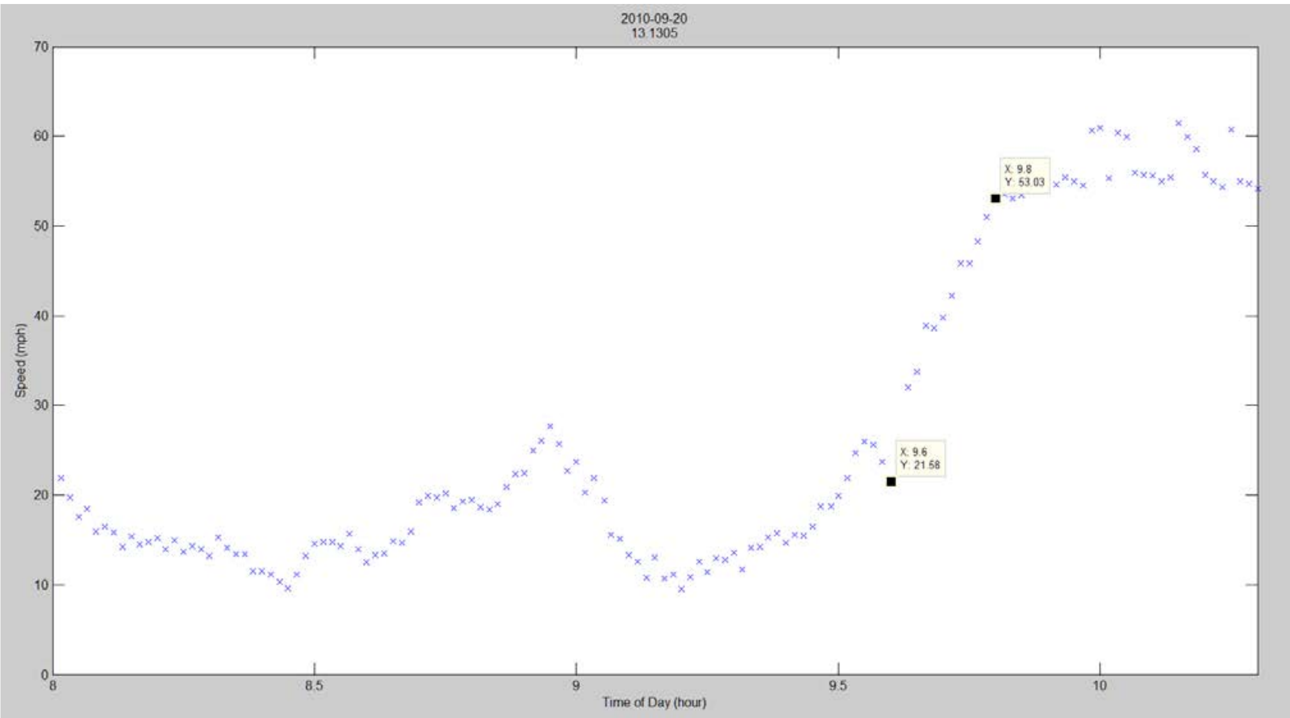
It is obvious that in order to apply our validation process, we will need a certain number of data points. The number of data points should be such that most of the features of the traffic can be studied. According to Nyquist-Shannon sampling theorem, the frequency of the data points we get from the reference state should be twice as much as the frequency of evolution of the traffic. This condition on the frequency combined with the penetration rate of the technology used for the references state will imply constraints on the flow of vehicles at the locations considered as candidates for the reference state implementation.

Here is an actual typical peak observed by radar 158997:



As we can see, there is congestion between approximately 7:30 and 10 AM. After 10 AM, we can observe that there are two main ranges of vehicles: those travelling at around 54mph and those travelling at around 60mph.

The following graph shows a zoom on the peak so that we can infer what the constraint on the resolution of our reference state measurement should be.



Let us assume that we want to be able to observe variations of speeds of 2 mph. The increase that can be measured thanks to the two data tips above shows that speed can evolve at a rate as high as 2.6mph/min, i.e., 2mph every 46 seconds. Consequently, according to the Nyquist-Shannon criterion, we need a frequency of measurement f_{needed} of 2 points every 46 seconds, which means 2.6 points/min or about **160 points/hr.**

We assume a penetration rate of the reference state technology equal to N out of 100 vehicles. Since the frequency of measurement we get is equal to $f = q * N$, we need to be at a location where the flow is at least $q_{\min} = \frac{f_{needed}}{N}$. In the case of the Bluetooth technology, N = 5%. Consequently, the flow at the location considered for the implementation of such a reference state should be greater than 3,200 vph.

Moreover, since the data feeds will be compared to the reference state, according to the Nyquist-Shannon theorem, the penetration rate of the reference state has to be at least twice as much as the penetration rate of the data feeds.

$$N_{ref\_state} \geq \max_{data\ feeds} (N_{data\_feed})$$

3.3 VARIABILITY BETWEEN DAYS OF WEEK AND TIME WINDOW CONSIDERATIONS

Another effect that our preliminary study enhanced is the periodicity of traffic between days of week. Indeed, we would find locations that have a high variability between days, but when looking at the variability between specific days of the weeks (ex: variability between Tuesdays or variability between Mondays) we found that the variability between these days was small. To refine our results we consequently introduced another criterion in the choice of our candidates that maximizes the variability between days of the weeks.

In addition to this, the study of the variability between the days of the weeks will give us good information with regard to the time window related to the reference state. As mentioned earlier, it could be that the technology used for our reference state imposes constraints on the time window during which the comparison can be made.

A basic index *Ivar* was used to characterize the variability between days. For each location we computed $\overline{d_{dayOfWeek}}$ and $\sigma_{dist}^{dayOfWeek}$ (where $dayOfWeek \in \{monday, tuesday, wednesday, thursday, friday\}$) which are respectively the mean of the L1 errors between the same days of the weeks and the standard deviation of the L1 errors between the same days of the weeks.

The following table summarizes the possible combinations and their meaning:

	$\boldsymbol{\sigma_{dist}^{dayOfWeek}}$ $\boldsymbol{high}$	$\boldsymbol{\sigma_{dist}^{dayOfWeek}}$ $\boldsymbol{low}$
$\boldsymbol{\sigma_{dist}}$ high	Variability between days is high Good candidate	Variability is high between different days of the week but the variability between Mondays, Tuesdays… is not high. Bad candidate
$\boldsymbol{\sigma_{dist}}$ low	X	Variability between days is low Bad candidate

The case of a low $\boldsymbol{\sigma_{dist}}$and a high $\boldsymbol{Ivar}$ is not possible because a high $\boldsymbol{Ivar}$ implies a high variability between days.

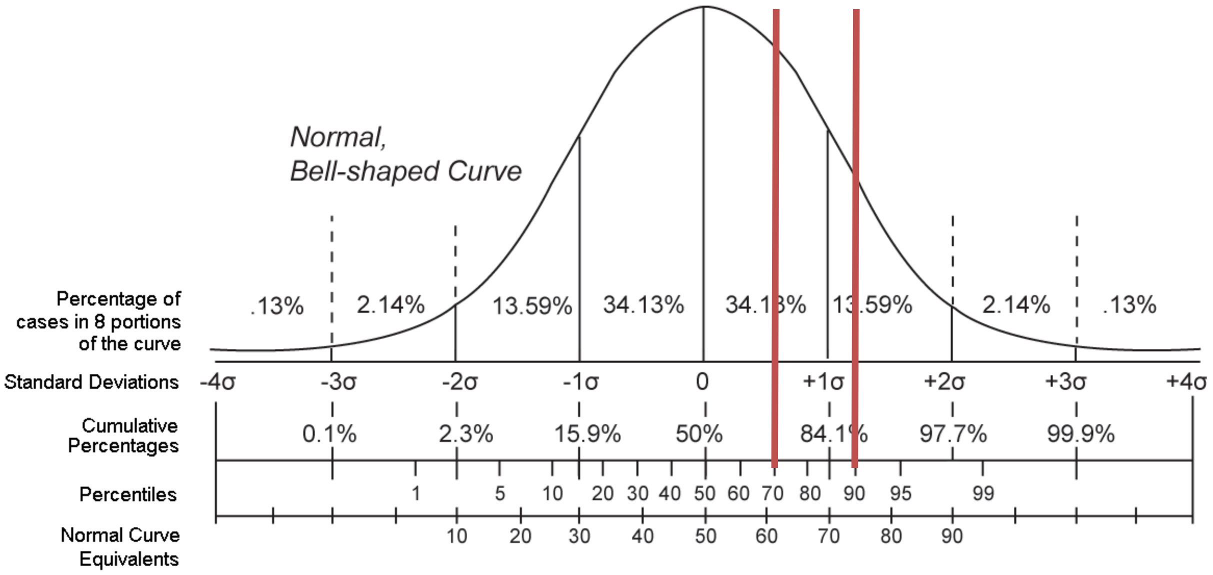
3.4 CANDIDATES: HIGHLY VARIABLE BUT STILL REPRESENTATIVE TRAFFIC

Let us consider a location for which each of the metrics defined above has a high value compared to other locations. Then at this location:

- Traffic evolves a lot during a day, i.e., the variability of traffic during a day is high *(σ is high)*
- Traffic is very different from one day to another *($\bar{d}$ and σ_{dist} are high)*

Thus, such a place can be considered as a place where the variability of traffic is high. Consequently, this place is a good candidate for the place where the reference state should be built.

Obviously, we do not want this location to be "an extreme" since it has to be representative. In order to take into account the necessity for representativeness, we will assume that such locations are between the 70th and 90th percentiles of the distribution of the values of the metrics.



(Wikipedia)

3.5 HIERARCHY OF CANDIDATES

Characteristics of the location considered	σ higher	σ between the 70th and 90th percentiles of the distribution	σ between the 50th and 70th percentiles of the distribution	σ lower
$\boldsymbol{\sigma_{dist}}$ higher	Not representative	Not representative	Not representative	Not representative
$\boldsymbol{\sigma_{dist}}$ high, i.e., between the 70th and 90th percentiles of the distribution	Not representative	Best candidates (both representative and highly variable traffic)	Secondary candidates	Not considered
$\boldsymbol{\sigma_{dist}}$ between the 50th and 70th percentiles of the distribution	Not representative	Secondary candidates	Not considered	Not considered
$\boldsymbol{\sigma_{dist}}$ lower	Not representative	Not considered	Not considered	Not considered

For a given location, if one of the two metrics studied has a too high value, this location is not considered as a candidate to build the reference state because it is not representative of traffic.

For a given location, if one of the two metrics studied has a too low value, this location is not considered since the variability of traffic cannot be considered to be high at this place.

Thus, as shown on the table above, three categories of locations remain: those for which one of the metrics is high and still representative ("high but not too high") and the other is "lower but not too low". Finally the best candidates are obviously the locations where the two metrics are high and not high enough to lose their feature of representativeness.

In addition to this, it is important to understand the distinctions between $\boldsymbol{\bar{d}}$ and $\boldsymbol{\sigma_{dist}}$. In order to do so, the following table categorizes locations depending on the order of magnitudes of the values obtained for $\boldsymbol{\bar{d}}$ and $\boldsymbol{\sigma_{dist}}$.

	$\boldsymbol{\bar{d}}$ high	$\boldsymbol{\bar{d}}$ low
$\boldsymbol{\sigma_{dist}}$ high	Good candidate + Traffic is difficult to predict + Traffic is very different from one day to another	+ Traffic is difficult to predict - Traffic does not vary a lot from one day to another
$\boldsymbol{\sigma_{dist}}$ low	- Traffic is not difficult to predict + Traffic can differ a lot from one day to another	Bad candidate - Traffic is not difficult to predict - Traffic is quite similar from one day to another

3.6 RECOMMENDATIONS FOR THE REFERENCE STATE

- Weekends should not be taken into account if it enables the study of more weekdays.

- The penetration of the technology used for the reference state should be at least twice as high as the penetration rate of the technologies used to derive the data feeds assessed. Indeed, this is necessary for the reference state to be legitimate in the assessment of the quality of data feeds.

- What is more, the penetration rate of the reference state directly implies a constraint on the minimum flow that has to be observed at the location chosen for the validation process. Thus, one should make sure that the relationships established earlier are verified.

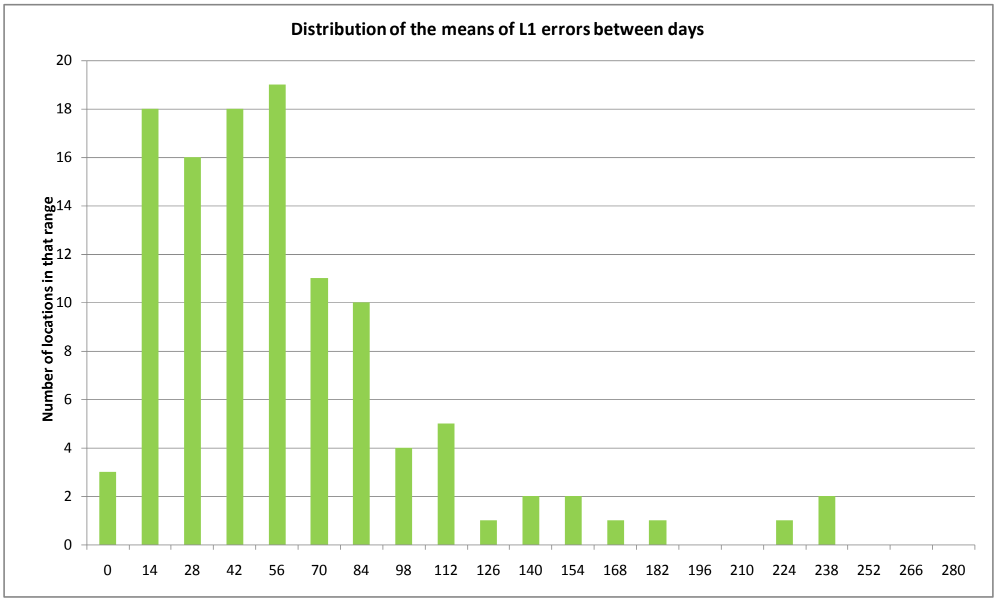
- Data should be collected in all speed ranges in order to have a complete data quality assessment. In addition, the assessment should be made on various profiles for the evolution of the speed (quick transitions from free flow to congestion and from congestion to free flow). To make sure a certain level of variability will be observed, the location has to be among the best candidates, as defined earlier. This ensures both variability during the day and variability between days.

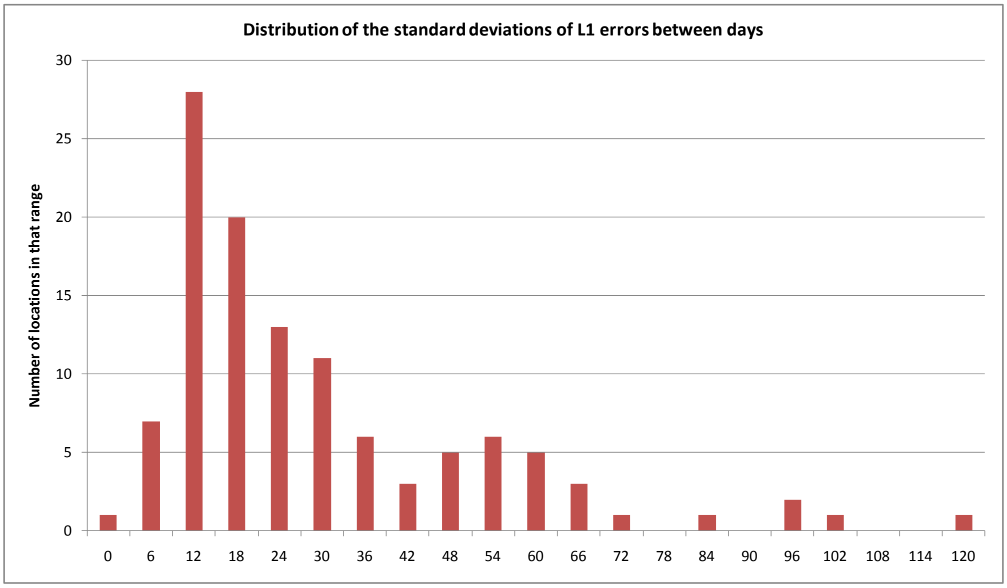
4 RESULTS

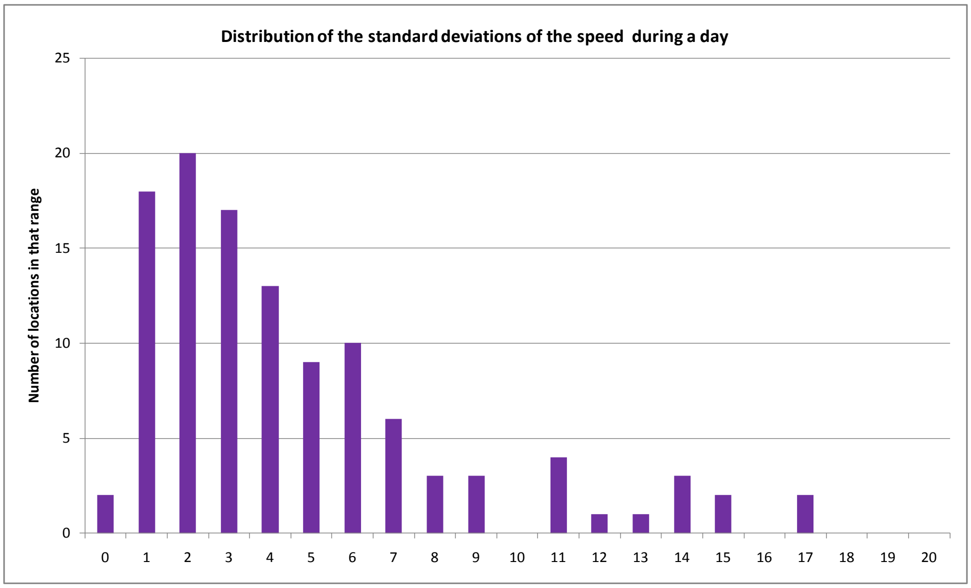
In total, 114 radars were studied.

4.1 DISTRIBUTION OBTAINED FOR THE METRICS

As we can see, the shape of these distributions does not look like a normal distribution. The right tail is surprisingly long since we have locations where the metric reaches values almost six times greater than the average value.

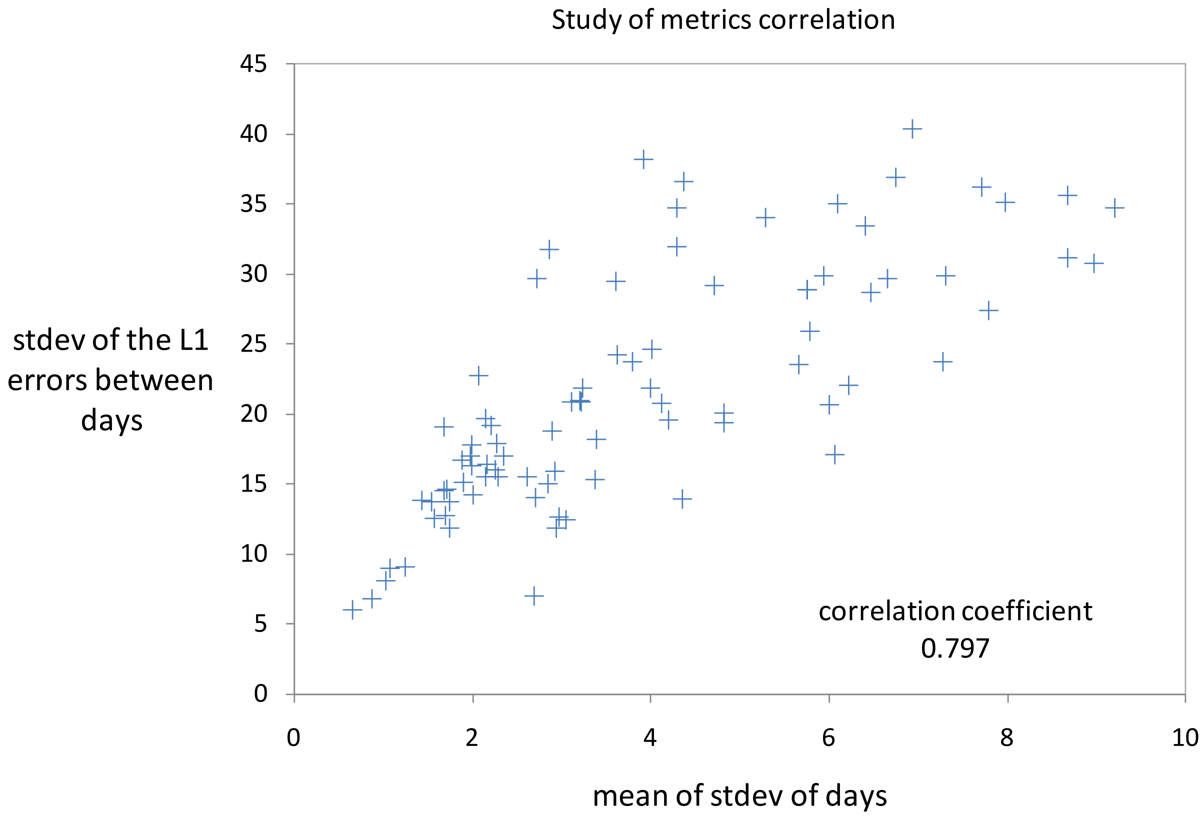





Distribution of the standard deviations of the speed during a day
Number of locations in that range
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

4.2 CORRELATION BETWEEN METRICS

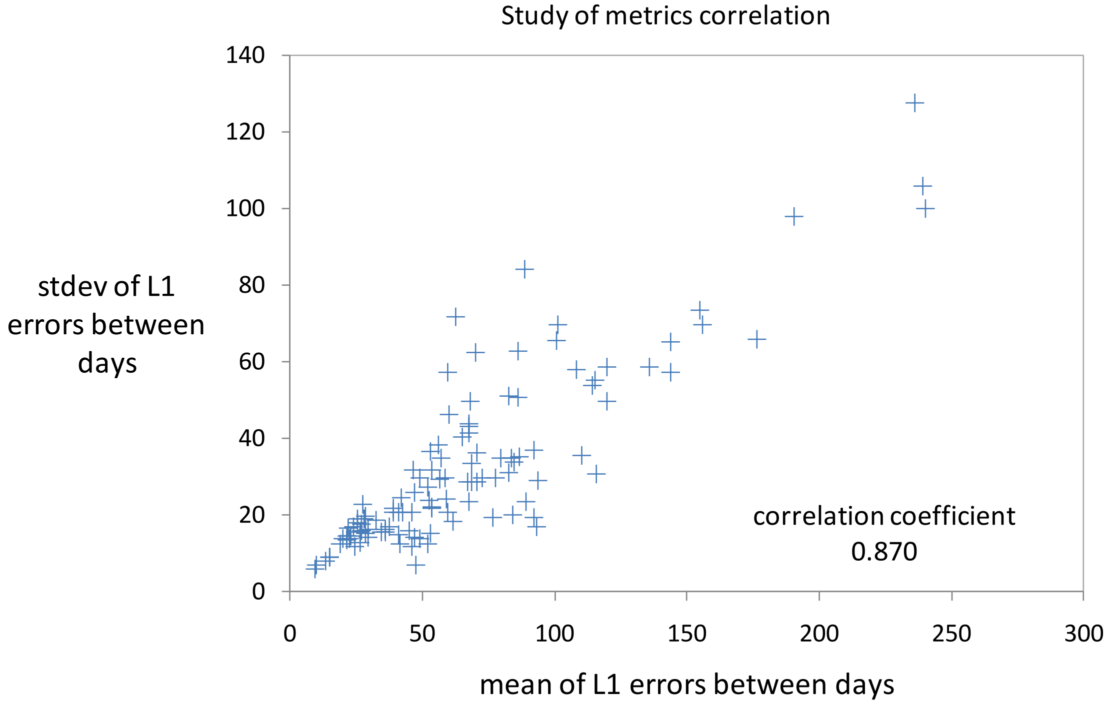
Scatter plots of are drawn with the data we get from the radar feed. Each point represents one radar. The correlation coefficient[7] is also computed.



This graph helps us to put into perspective the table comparing the locations with different $\boldsymbol{\sigma_{dist}}$and $\overline{\boldsymbol{d}}$. The strong correlation between these two metrics shows that we would have either good candidates or bad candidates. Locations with only one high value for these two metrics are very unlikely.

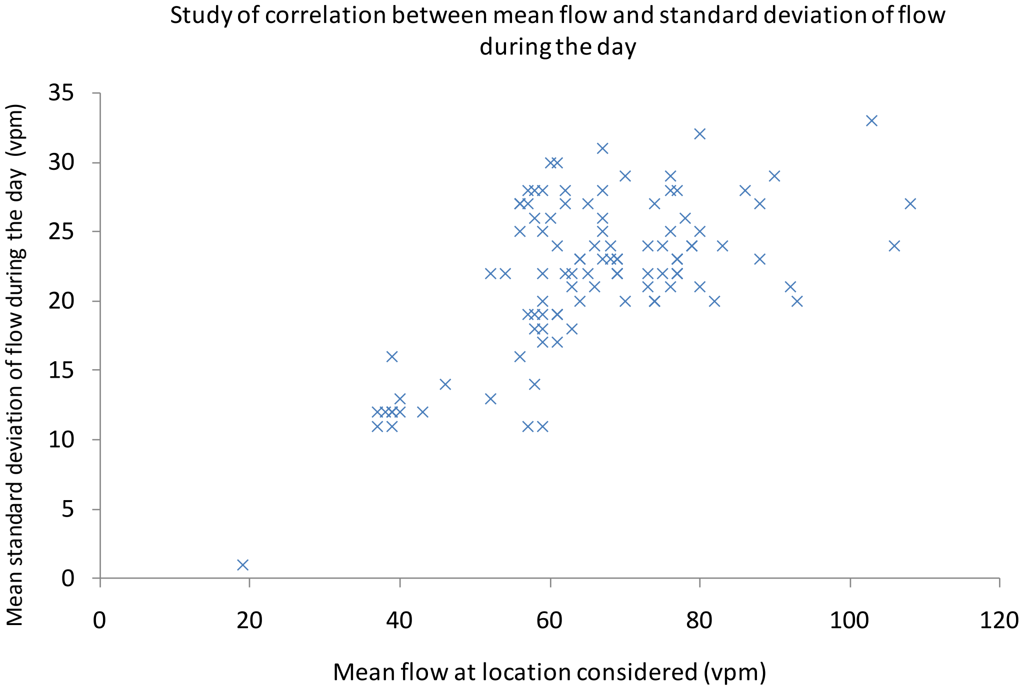
---

[7] Note: the correlation coefficient between X and Y is given by the following formula: $(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$.



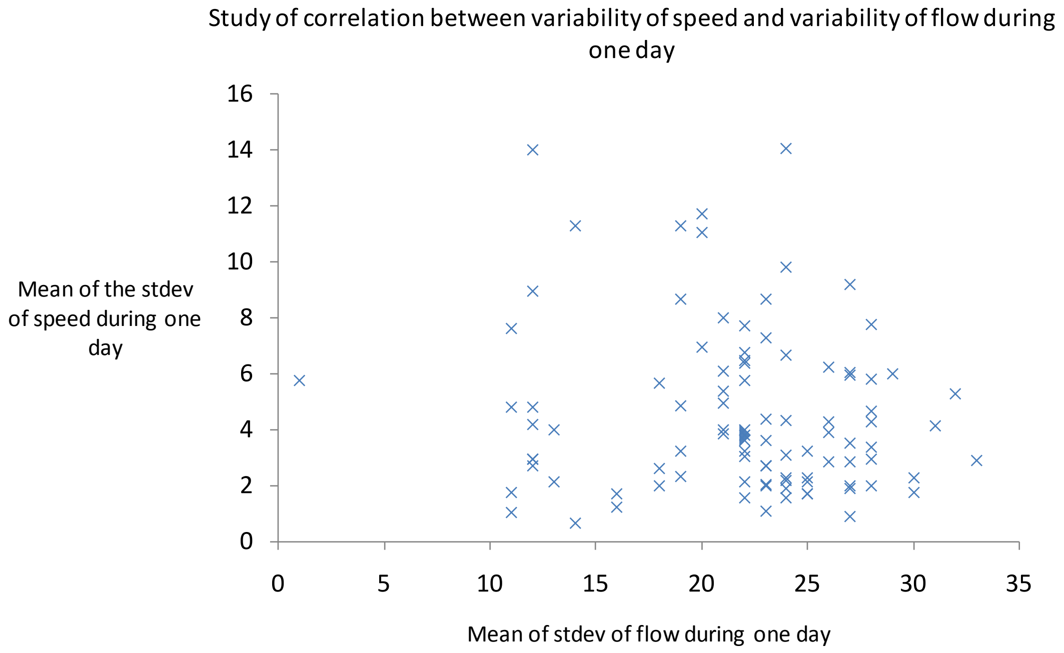
This graph leads to the following interesting conclusion: Where traffic can evolve a lot during the day (σ is high), its evolution is very different from one day to another as well (σ_{dist} is high).

4.3 STUDY OF FLOW



As we can see on the graph above, the mean flow at radar locations does not go under 30 vpm except for one location. What is more, the values standard deviation of flow during a day at the radar locations ensure that the flow is unlikely to be less than 30 vpm. This corresponds to 1,800 vph.

The correlation between the variability of flow and variability of speed during one day is not very strong.



This must be due to the fact that traffic theory shows that there are two flows possible corresponding to a speed.

4.4 RESULTS FOR THE BAY AREA

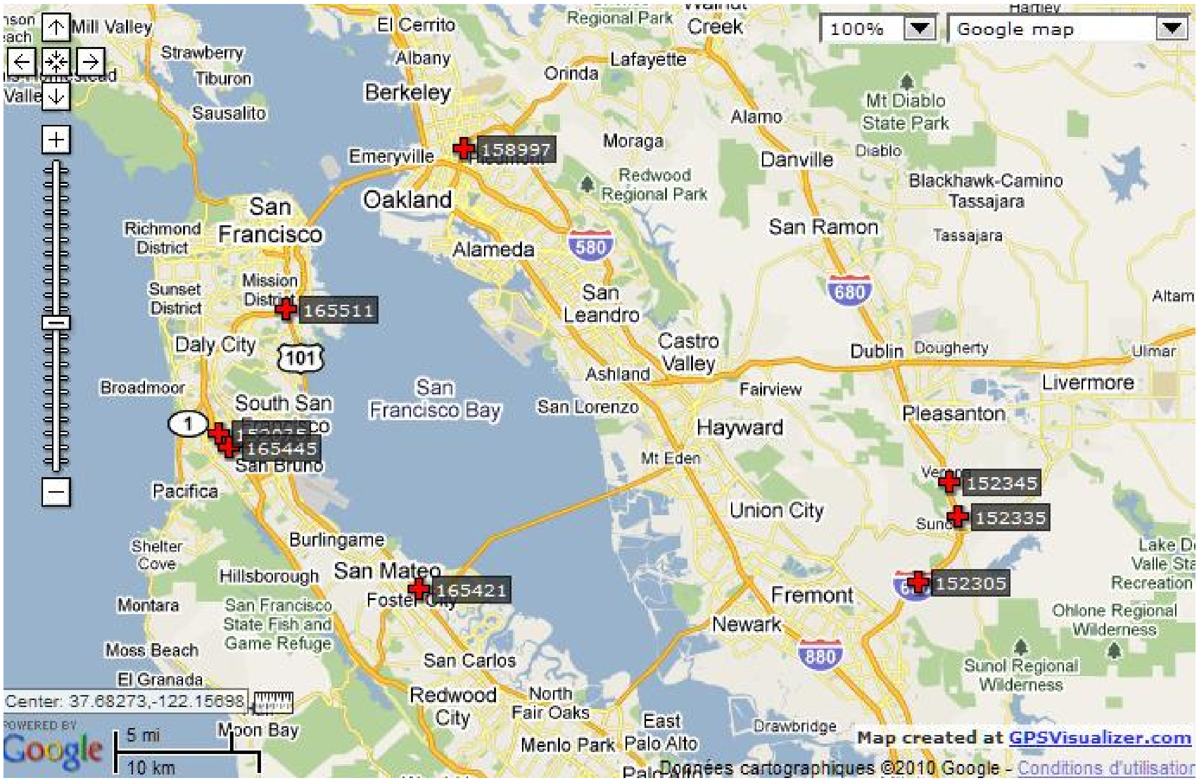
As explained in the methodology, for the locations that are between the 70th and 90th percentiles for only one metric, we will consider this place as a LRHVRT[8] only if it is between the 50th and 90th percentiles for the other metrics.

Numbers of locations found	σ higher	σ between the 70th and 90th percentiles of the distribution	σ between the 50th and 70th percentiles of the distribution	σ lower
$\boldsymbol{\sigma_{dist}}$ higher	9	1	1	1
$\boldsymbol{\sigma_{dist}}$ high i.e. between the 70th and 90th percentiles of the distribution	3	8	6	5
$\boldsymbol{\sigma_{dist}}$ between the 50th and 70th percentiles of the distribution	0	10	9	4
$\boldsymbol{\sigma_{dist}}$ lower	0	4	5	47

---

[8] Location Representative of Highly Variable Traffic

We found 8 locations which are between the 70th and 90th percentiles for both metrics (σ and σ_{dist}). In addition to these 8 "best candidates", we get 16 "secondary candidates".

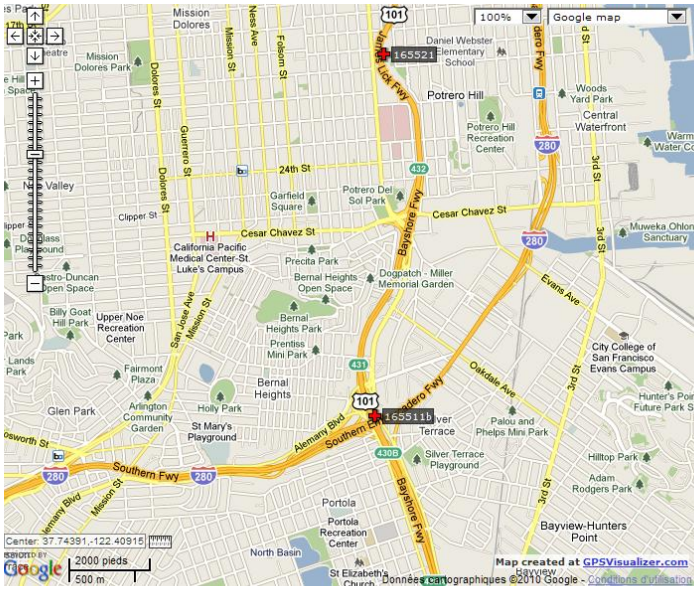


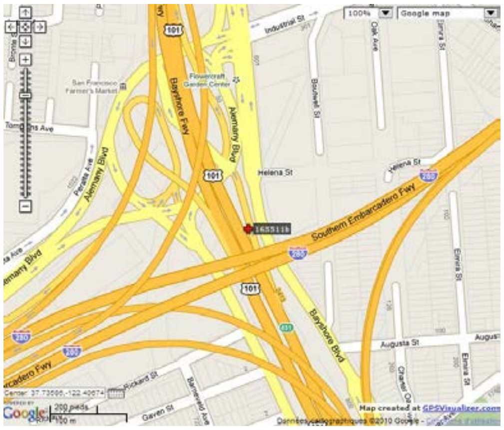
Using the GPS visualizer (Google GPS visualizer, 2010), we can draw the locations of the radars on a map.

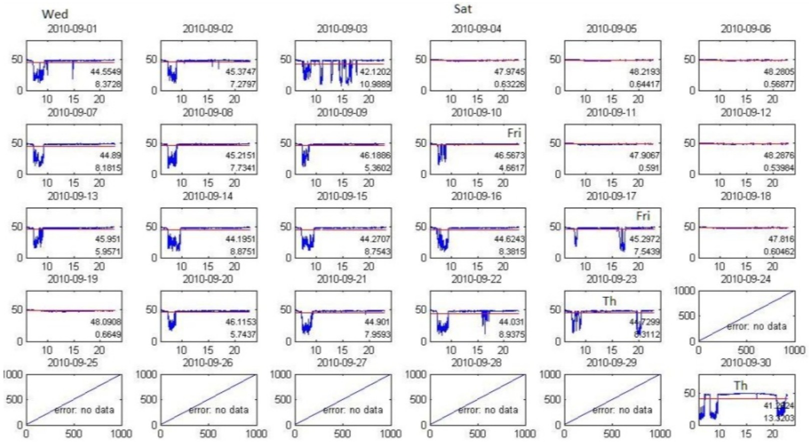
You can compare this map to the map showing the 114 working radars of the Bay Area (see section 6). The 8 locations shown are the best candidates, as defined earlier.

4.4.1 HIGHLY VARIABLE AND HIGHLY REPRESENTATIVE LOCATIONS: BEST CANDIDATES FOUND

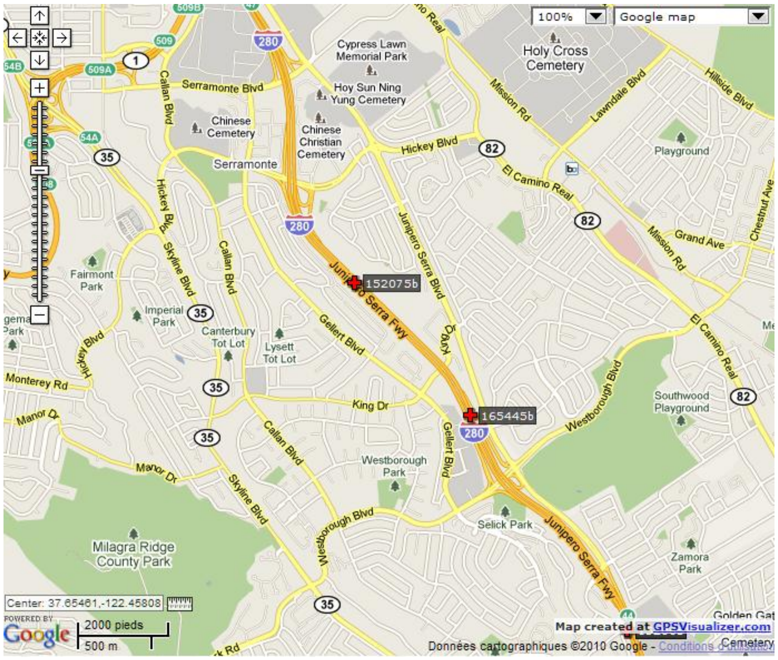
165511: Notice that this location I rights at the intersection of two major highways: US-101 and I-280. It is, by far, the closest radar to this intersection.

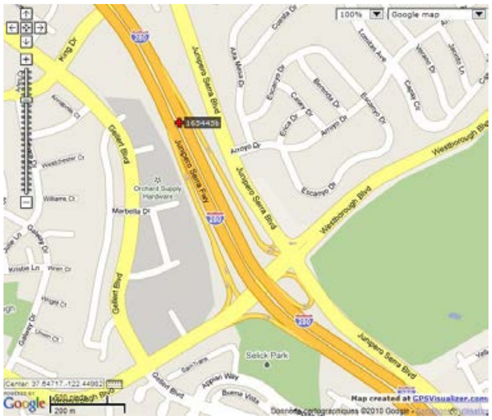


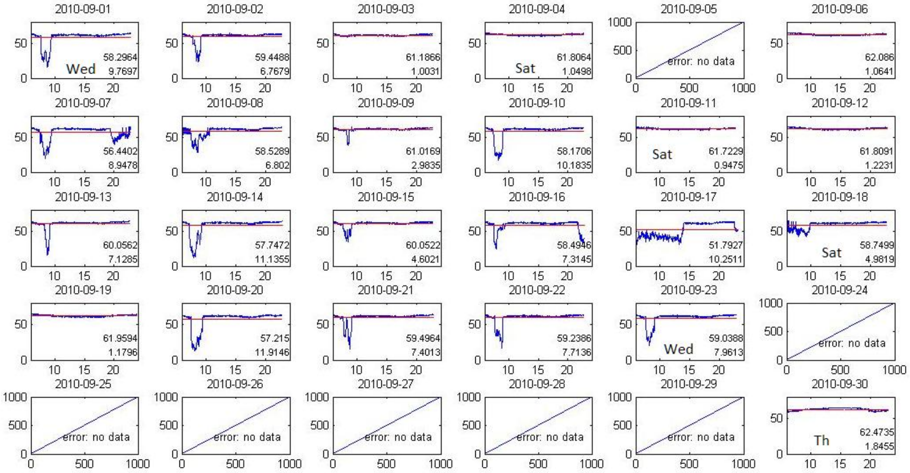




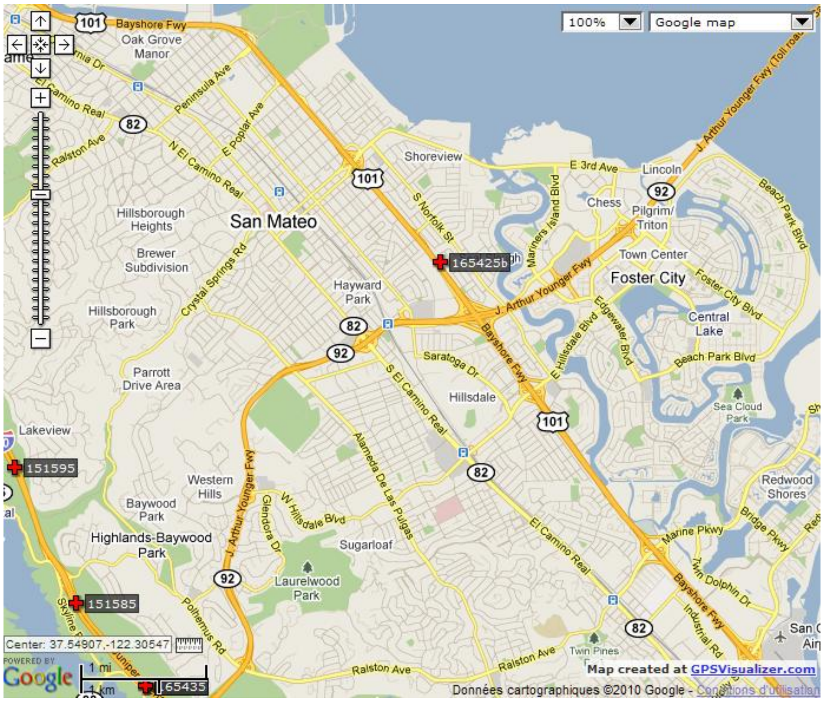
165445 and 152075: These two locations being pretty close to each other, we can reasonably assume that the high variability of traffic of 152075 results from 165445. This location is close to a major entry on highway I-280. It is a very residential area. This radar is the closest one to the intersection between 280 and Westborough Rd.

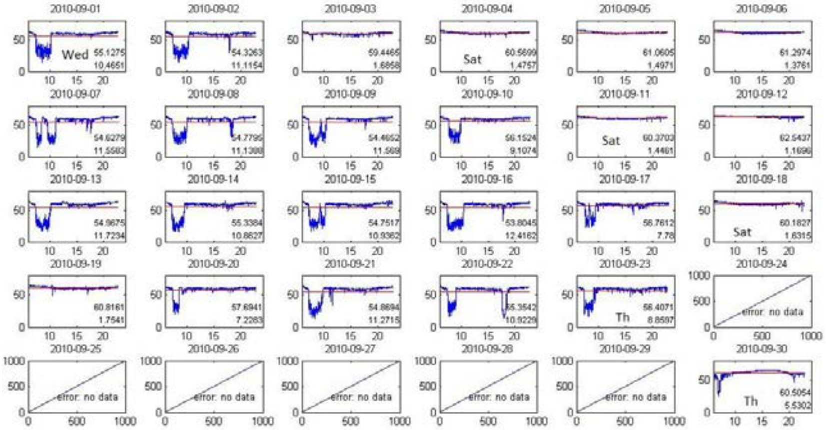




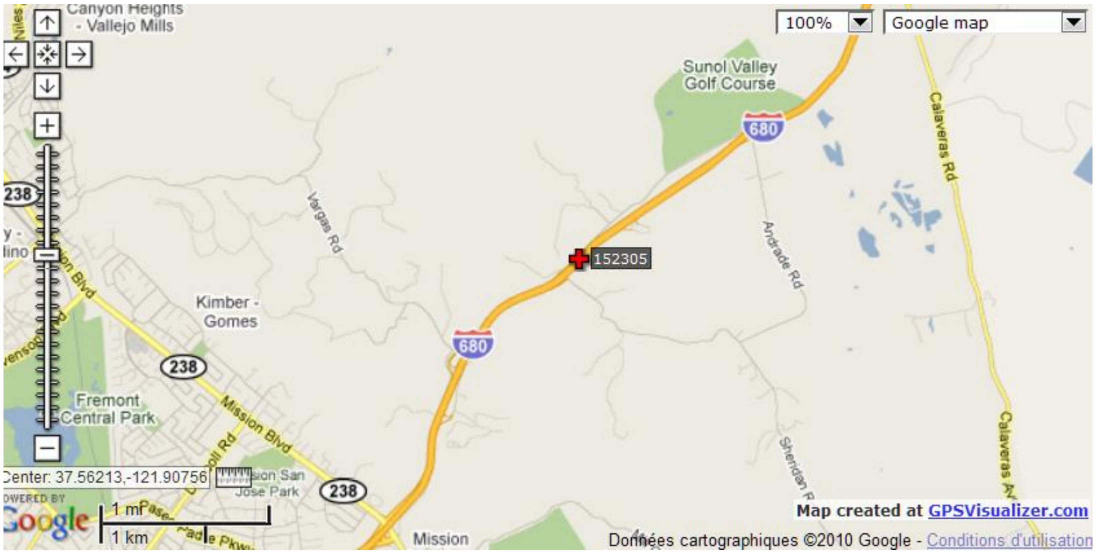


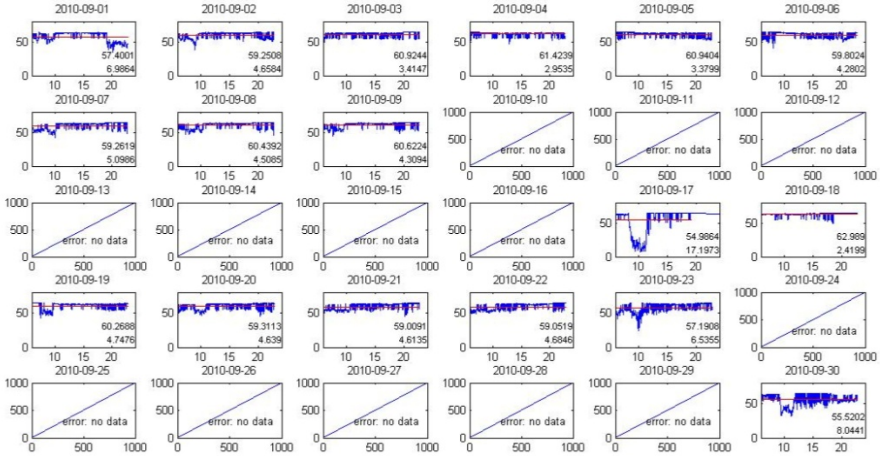
165425: Once again, this location is located very close to a major intersection of two highways: 92 and 101. It is a very residential area. This radar is the only one in this area.





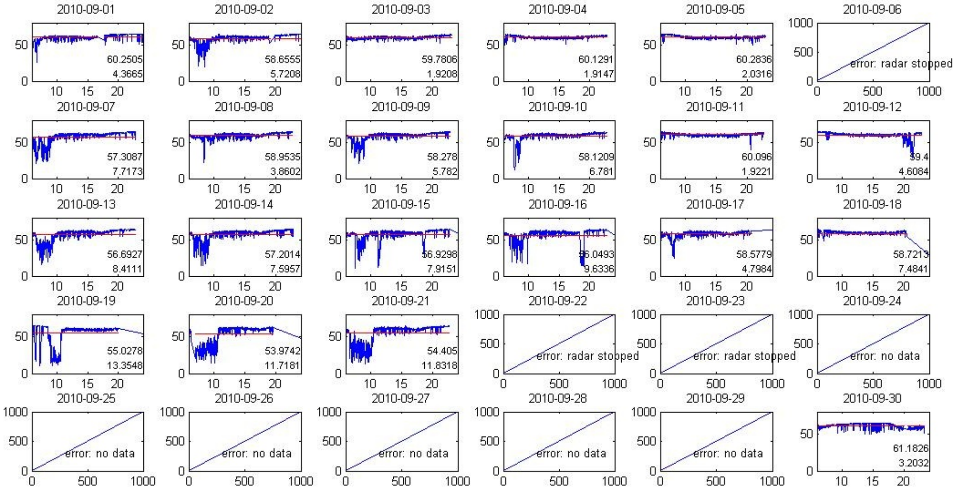
152305:


100%
Google map
Canyon Heights - Vallejo Mills
Niles
Sunol Valley Golf Course
680
Calaveras Rd
238
Vargas Rd
Andrade Rd
152305
Kimber - Gomes
680
238
Fremont Central Park
Mission Blvd
Driscoll Rd
Sheridan Rd
Calaveras Av
Center: 37.56213,-121.90756
Mission San Jose Park
238
POWERED BY Google
1 mi
1 km
Mission
Map created at GPSVisualizer.com
Données cartographiques ©2010 Google - Conditions d'utilisation

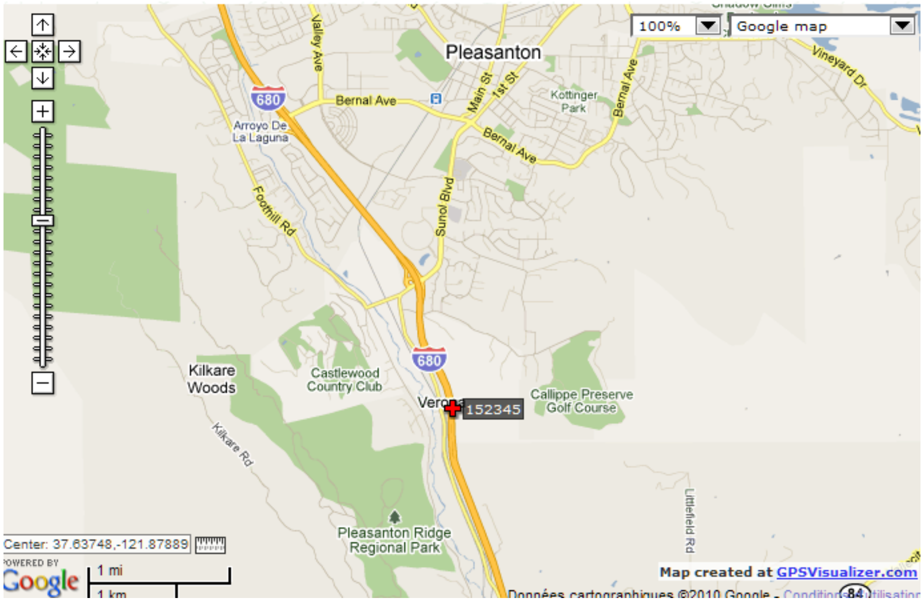

2010-09-01
57.4001
6.9864
2010-09-02
59.2508
4.6584
2010-09-03
60.9244
3.4147
2010-09-04
61.4239
2.9535
2010-09-05
60.9404
3.3799
2010-09-06
59.8024
4.2802
2010-09-07
59.2619
5.0986
2010-09-08
60.4392
4.5085
2010-09-09
60.6224
4.3094
2010-09-10
error: no data
2010-09-11
error: no data
2010-09-12
error: no data
2010-09-13
error: no data
2010-09-14
error: no data
2010-09-15
error: no data
2010-09-16
error: no data
2010-09-17
54.9864
17.1973
2010-09-18
62.989
2.4199
2010-09-19
60.2688
4.7476
2010-09-20
59.3113
4.639
2010-09-21
59.0091
4.6135
2010-09-22
59.0519
4.6846
2010-09-23
57.1908
6.5355
2010-09-24
error: no data
2010-09-25
error: no data
2010-09-26
error: no data
2010-09-27
error: no data
2010-09-28
error: no data
2010-09-29
error: no data
2010-09-30
55.5202
8.0441
0
50
10
15
20
500
1000

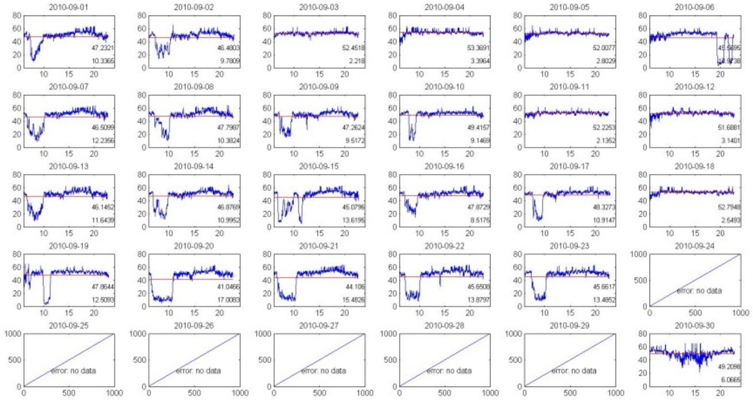
152335:


100%
Google map
Carver Ln
Kilkare Rd
Foothill Rd
152335
Alpha Ln
Little Valley Rd
84
680
State Hwy 84
Vallecitos Rd
84
21A
Sunol
84
Main St
84
Paloma Rd
21B
Scotts Corner
Center: 37.59778,-121.86919
POWERED BY Google
2000 pieds
500 m
Map created at GPSVisualizer.com
Données cartographiques ©2010 Google - Conditions d'utilisation

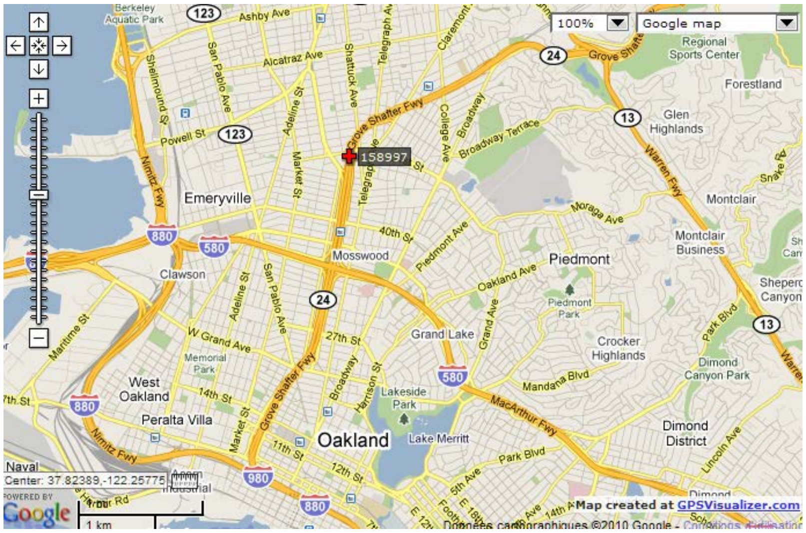

2010-09-01
60.2505
4.3665
2010-09-02
58.6555
5.7208
2010-09-03
59.7806
1.9208
2010-09-04
60.1291
1.9147
2010-09-05
60.2836
2.0316
2010-09-06
error: radar stopped
2010-09-07
57.3087
7.7173
2010-09-08
58.9535
3.8602
2010-09-09
58.278
5.782
2010-09-10
58.1209
6.781
2010-09-11
60.096
1.9221
2010-09-12
59.4
4.6084
2010-09-13
56.6927
8.4111
2010-09-14
57.2014
7.5957
2010-09-15
56.9298
7.9151
2010-09-16
56.0493
9.6336
2010-09-17
58.5779
4.7984
2010-09-18
58.7213
7.4841
2010-09-19
55.0278
13.3548
2010-09-20
53.9742
11.7181
2010-09-21
54.405
11.8318
2010-09-22
error: radar stopped
2010-09-23
error: radar stopped
2010-09-24
error: no data
2010-09-25
error: no data
2010-09-26
error: no data
2010-09-27
error: no data
2010-09-28
error: no data
2010-09-29
error: no data
2010-09-30
61.1826
3.2032
0
50
1000
500
10
15
20

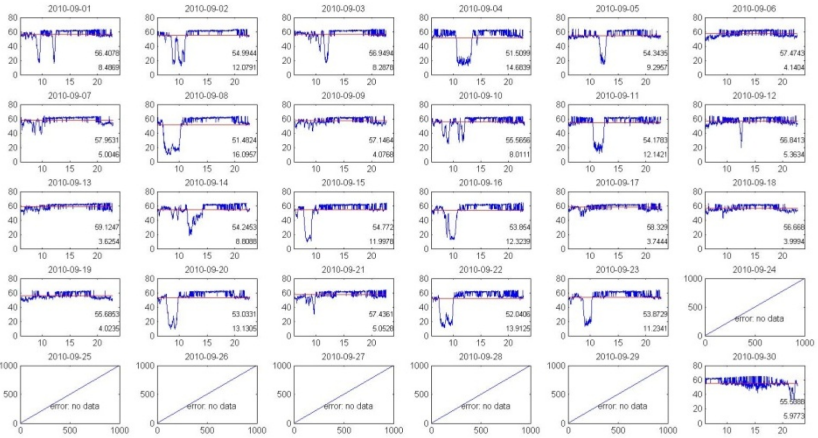
152345:


100%
Google map
Pleasanton
680
Bernal Ave
Valley Ave
Main St
1st St
Kottinger Park
Arroyo De La Laguna
Foothill Rd
Sunol Blvd
Vineyard Dr
Kilkare Woods
Castlewood Country Club
Verona
152345
Callippe Preserve Golf Course
Kilkare Rd
Pleasanton Ridge Regional Park
Littlefield Rd
84
Center: 37.63748,-121.87889
POWERED BY Google
1 mi
1 km
Map created at GPSVisualizer.com
Données cartographiques ©2010 Google - Conditions d'utilisation


2010-09-01
47.2321
10.3365
2010-09-02
46.4803
9.7809
2010-09-03
52.4518
2.218
2010-09-04
53.3691
3.3964
2010-09-05
52.0077
2.8029
2010-09-06
45.5695
18.9738
2010-09-07
46.5099
12.2356
2010-09-08
47.7987
10.3824
2010-09-09
47.2624
9.5172
2010-09-10
49.4157
9.1469
2010-09-11
52.2253
2.1352
2010-09-12
51.6881
3.1401
2010-09-13
46.1452
11.6439
2010-09-14
46.8769
10.9952
2010-09-15
45.0796
13.6195
2010-09-16
47.8729
8.5175
2010-09-17
48.3273
10.9147
2010-09-18
52.7948
2.5493
2010-09-19
47.8644
12.5093
2010-09-20
41.0466
17.0083
2010-09-21
44.106
15.4826
2010-09-22
45.6508
13.8797
2010-09-23
45.6617
13.4852
2010-09-24
error: no data
2010-09-25
error: no data
2010-09-26
error: no data
2010-09-27
error: no data
2010-09-28
error: no data
2010-09-29
error: no data
2010-09-30
49.2098
6.0665
0
20
40
60
80
10
15
20
500
1000

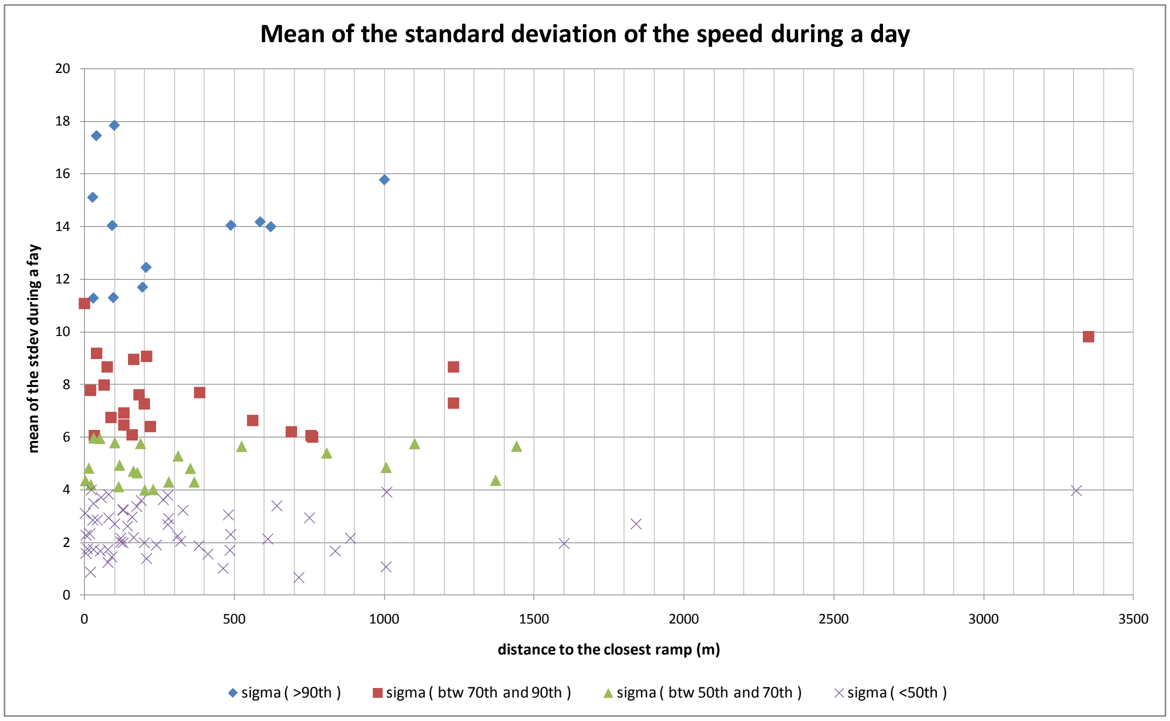
158997:



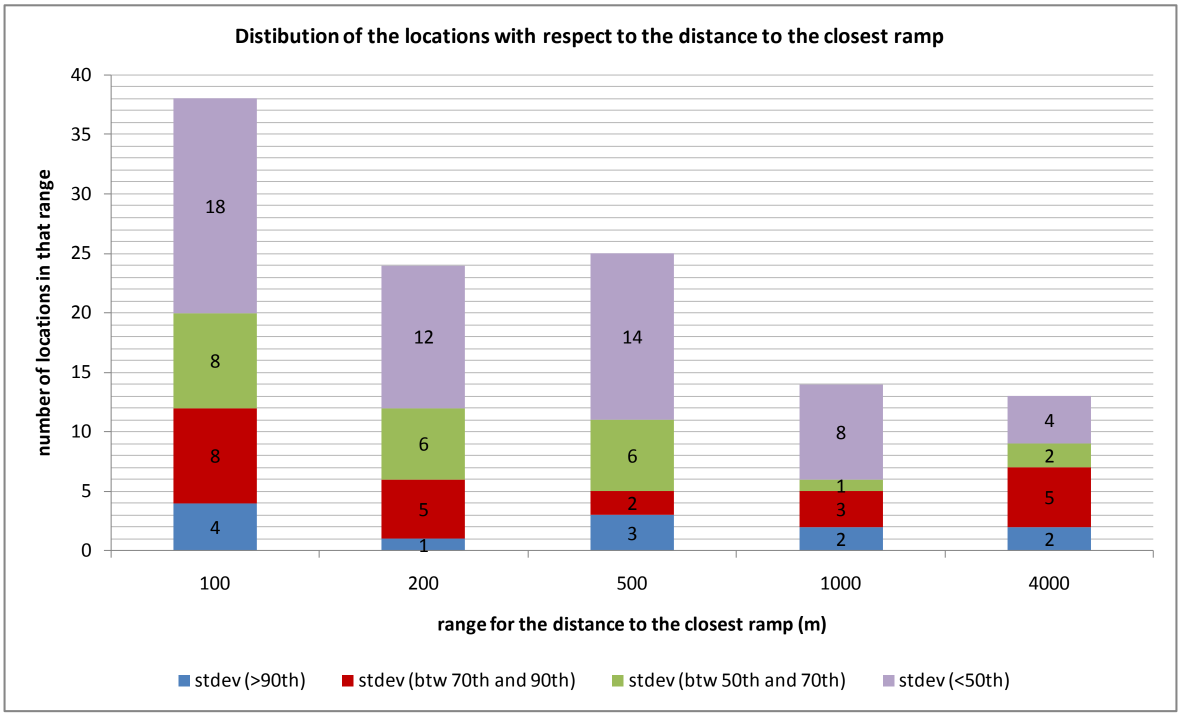


4.4.2 EVOLUTION OF THE VARIABILITY WITH THE DISTANCE TO THE CLOSEST RAMP

The distance to the closest ramp was measured for each radar using Google Maps (Google, 2010). Then the repartition of the locations depending on their distance to the closest ramp and depending on the variability of traffic at this place was studied. In order to do so, the following scatter plots were drawn for the three metrics of variability defined earlier.



The following histogram synthesizes the results:



4.5 CONSIDERATIONS WITH REGARD TO THE TIME WINDOW

The table below shows that the there is no strong evidence that speed would vary more on a particular day of the week. Indeed, the average standard deviations of the speed during a day do not show significant differences with respect to the day of week considered. However, the second row shows that the variability between traffic evolution of Mondays, Tuesdays, and Wednesdays is higher than the variability of traffic between Wednesdays and Thursdays.

Day of week	**Monday**		**Tuesday**		**Wednesday**		**Thursday**		**Friday**
Average standard devitation of the speed during one day	**4,6**		**5,3**		**5,5**		**5,2**		**4,9**
Average L1 errors between a particular day of week	**80**		**78**		**56**		**55**		**83**

As explained earlier, we would like to compare the data feeds with the reference state on various profiles for the evolution of traffic. We would like to apply the validation process for all ranges of speeds; moreover, the latency can be assessed when the traffic varies enough during the day. The following table shows the values of the metrics introduced in this study for the best candidates for the implementation of the reference state. They have to be taken into account depending on the constraints imposed by the technology chosen.

Radar id	**mean of stdev of days**	**stdev of the L1 errors between days**	**Monday**		**Tuesday**		**Wednesday**		**Thursday**		**Friday**
	variability of speed during one day	*variability between days*	*mean of the L1 erros*	*mean of stdev of days*	*mean of the L1 erros*	*mean of stdev of days*	*mean of the L1 erros*	*mean of stdev of days*	*mean of the L1 erros*	*mean of stdev of days*	*mean of the L1 errors*
	mph	*miles*									
158997	9	36	112	7	109	6	106	13	89	10	105
152075	8	36	98	7	95	9	44	9	49	8	90
152335	7	37	145	8	104	8	81	5	78	7	70
165445	7	40	66	5	79	7	36	7	45	6	120

For instance, if the reference state can only be implemented for one day, then we would try to maximize only the variability during the day. Then, radar 152345 should be selected, and Monday and Wednesday would be the favorite time windows.

If the reference state can be implemented for several days, the time window should be chosen so that the L1 errors between days are maximized as well as the variability during the day. Thus, radar 152305 may not be a good choice. Radar 165421 would be the best candidate. Radar 152345 could be considered as well.

5 CONCLUSIONS

Variability of traffic is a necessary feature for the location where the reference state will be implemented.

There is a strong correlation between the variability during a day and the variability between days. Consequently, one could expect that if traffic varies a lot during one day, various profiles for the speed evolution during a day will be observed.

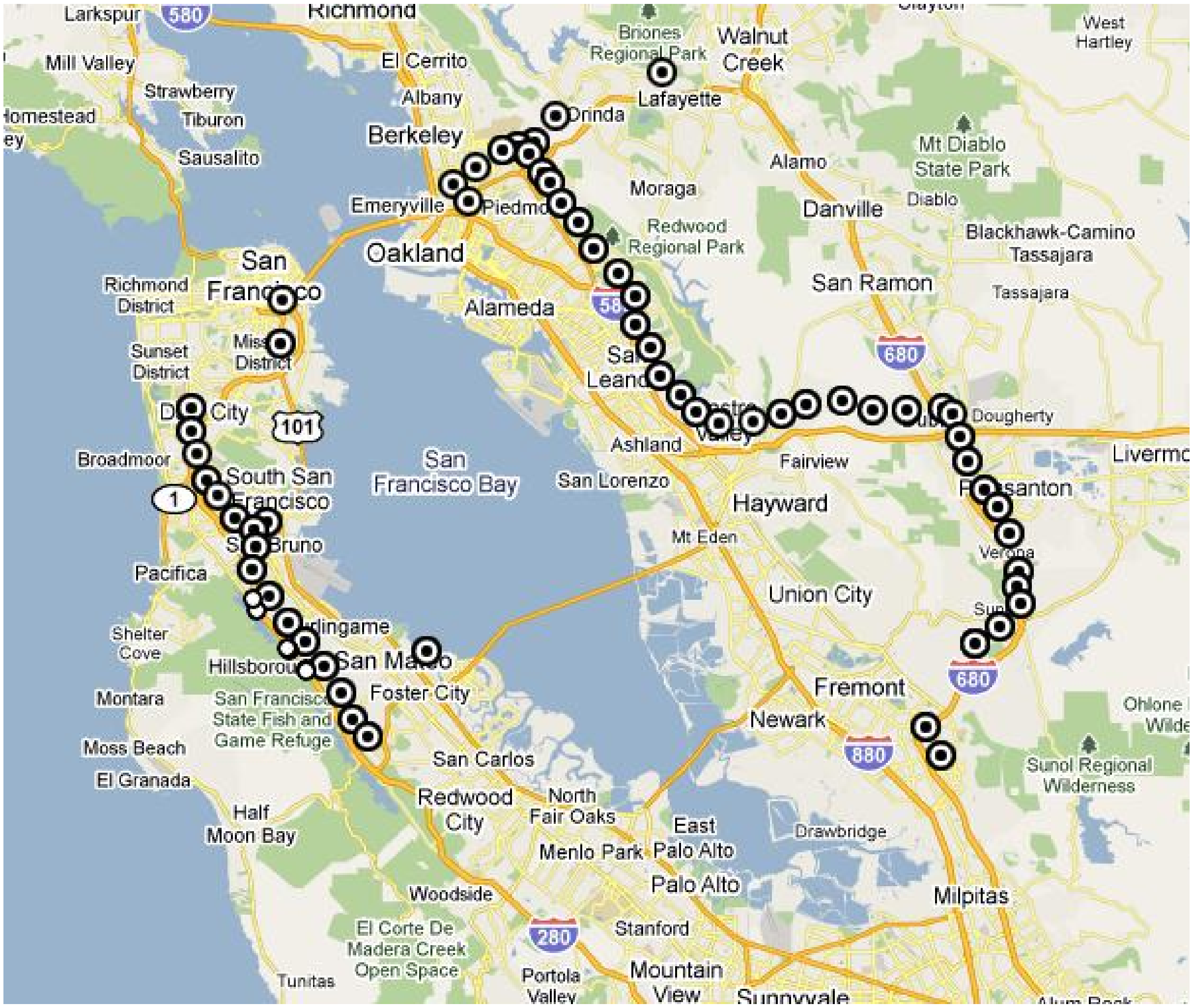
In order to abstract some necessary characteristics for the space window relevant to the building of our reference state, we studied the correlation between the distance to the closest ramp and the variability of traffic at the location considered. No convincing feature can be abstracted from this study. Consequently, if available, sources of data of a different nature should be used to study the variability of traffic in the area considered. Otherwise, empirical or historical observations should be used.

The technology used for the reference state is directly linked with the choice of the relevant location for its implementation. The higher the penetration rate, the lower the flow at the location considered has to be. What is more, the metrics we defined in this study help take into account the constraints imposed by the technology on the time window with regard to the implementation of the validation process.

6 RADAR LOCATIONS

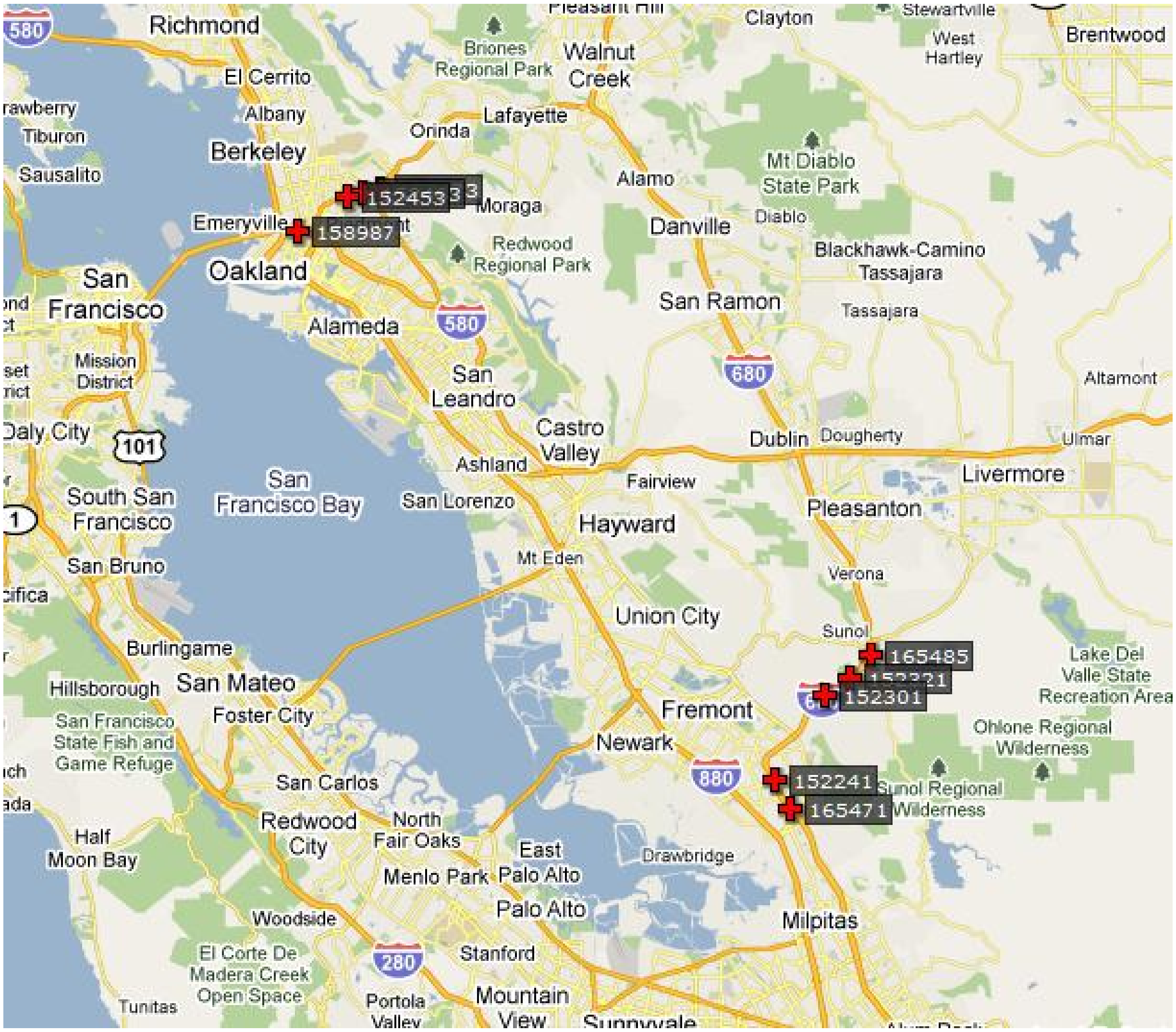
6.1 LOCATIONS OF ALL THE RADARS

On this chart, each circle represents a radar in our database that reports reliable values. In total, 114 radars are on this chart.

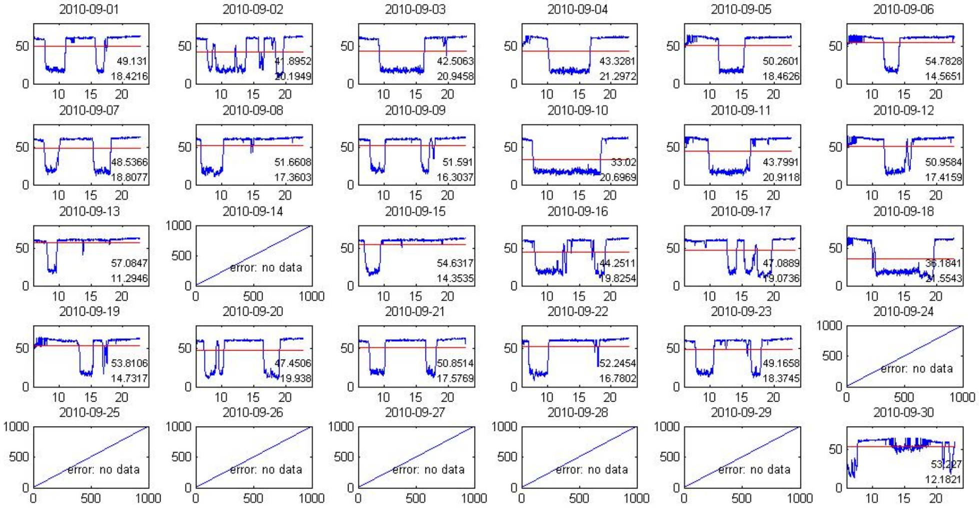


6.2 LOCATIONS WITH EXTREME VALUES

The following chart shows the locations which have a variability so high that they are not included in the 90th percentiles of the distributions of both σ and σ_{dist}.

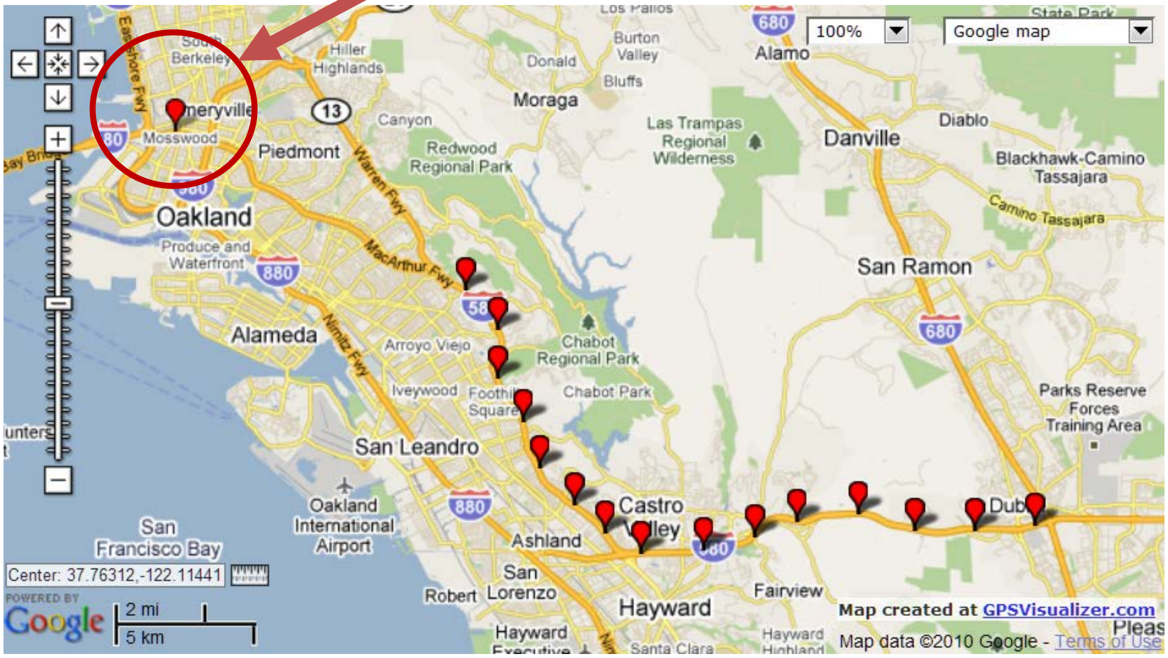


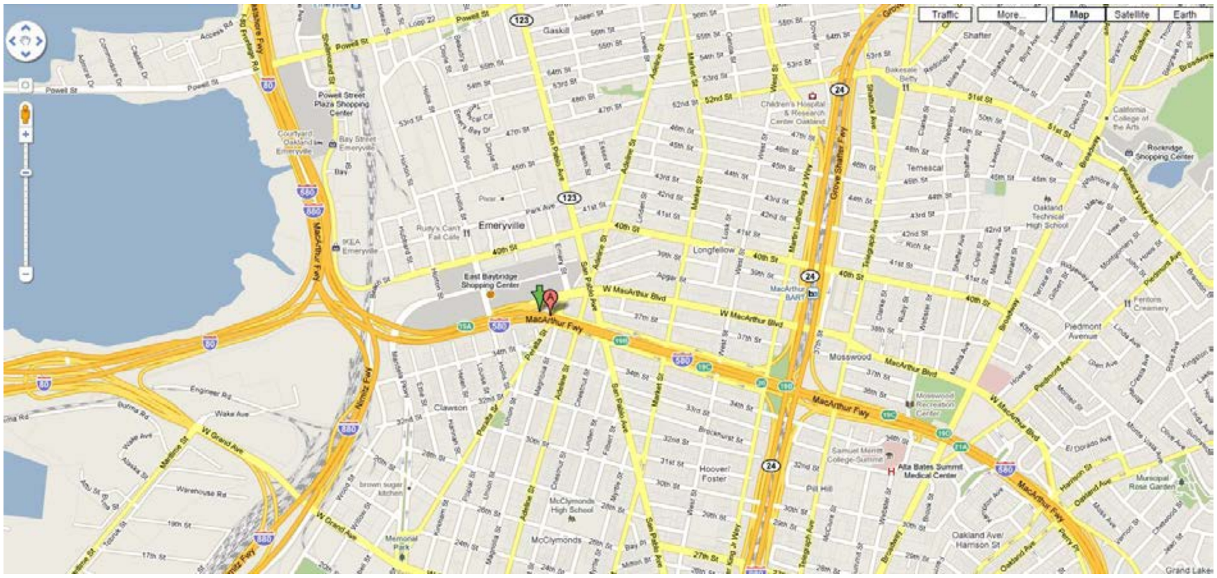
On highway I-580, one location presents the highest values for σ and σ_{dist} (radar_id: 158987). Its speed has been drawn for each day of the month of September 2010. The variability during the day and between days is indeed blatant.

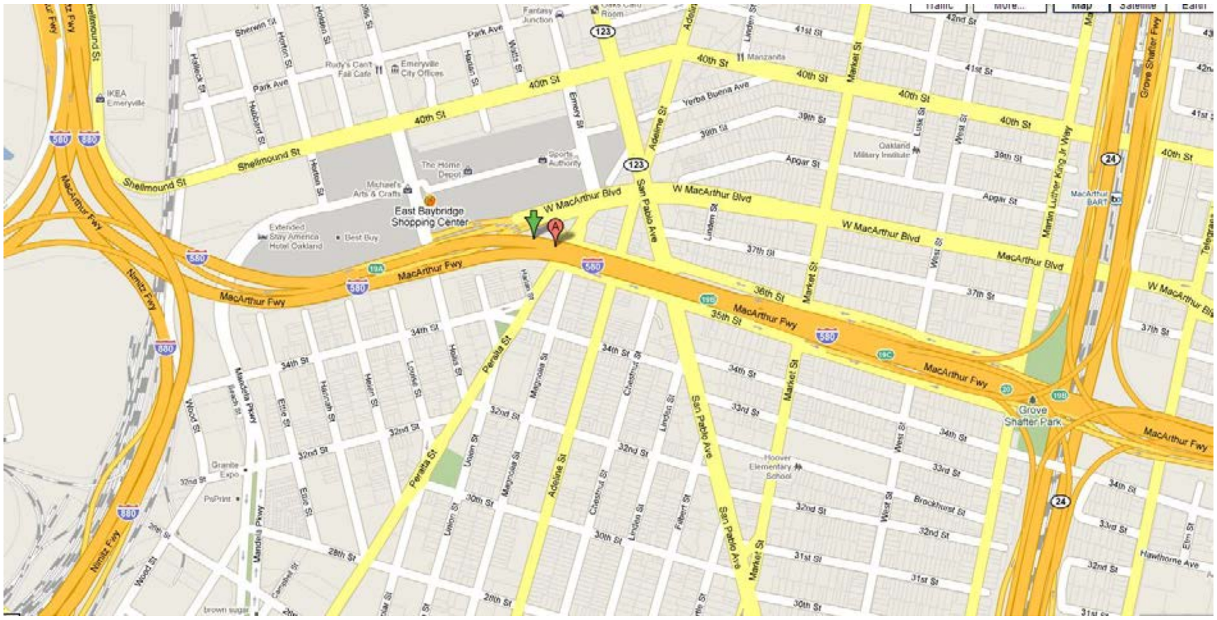


Radar locations on I-580:

Extreme variability







This radar is located on 580 in Emeryville almost at the crossroads between 880, 580, 80, 980 and 24, close to the entrance on the Bay Bridge.