Hybrid Data Implementation: Final Report for Task Number 3643

Khan, Sakib Mahmud, PhD
Fournier, Nicholas, PhD
Mauch, Michael, PhD
et al.

Publication Date
2020-12-01

PARTNERS FOR ADVANCED TRANSPORTATION TECHNOLOGY

INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

HYBRID DATA IMPLEMENTATION

FINAL REPORT FOR TASK NUMBER 3643

Prepared by:

Sakib Mahmud Khan, Ph.D., Post-Doctoral Scholar
Nicholas Fournier, Ph.D., Post-Doctoral Scholar
Michael Mauch, Ph.D., Research Engineer
Anthony D Patire, Ph.D., Research and Development Engineer
Alex Skabardonis, Ph.D., Professor In-Residence

PATH Research Report

CALIFORNIA PATH

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California. This report does not constitute a standard, specification, or regulation.

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
409a McLaughlin Hall, Berkeley, CA 94720-1720
Phone: (510) 642-5478 Fax: (510) 642-0910 http://www.path.berkeley.edu

UCB-ITS-PRR-2020-02 February 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AADT	Annual Average Daily Traffic
ATMS	Advanced Traffic Management System
C-GASM	Confined Generalized Adaptive Smoothing Method
Capsnet	Capsule Neural Network
CNN	Convolutional Neural Network
FATV	Fully Accounted Traffic Volume
GASM	Generalized Adaptive Smoothing Method
GPS	Global Positioning System
HOV	High Occupancy Vehicle
HOT	High Occupancy Toll
ICM	Integrated Corridor Management
MPR	Mobility Performance Report
OEM	Original Equipment Manufacturer
PeMS	Performance Measurement System
RNN	Recurrent Neural Network
SGT	Simulated Ground Truth
TMC	Traffic Management Center
VDS	Vehicle Detector Station
VHD	Vehicle Hours of Delay
VHT	Vehicle Hours Traveled
VMT	Vehicle Miles Traveled

LIST OF NOMENCLATURE

Cell: The term *cell* is used to denote the smallest domain of analysis in the algorithms presented here. A cell is a small length of freeway used to perform a fine level of analysis to narrow down areas of congestion. (Figure 1-1)

Data Conflation: *Conflation* is generally the projection of data from certain points on one map to other desired points (corresponding points) on a different map.

Data Fusion: The term *fusion* indicates the final integration of flow (from a traditional source) and travel time data (from a third-party).

Link: A *link* refers to a length of freeway for which travel time data is available from a third-party vendor. (Figure 1-1)

Section: The term *section* refers to a length of roadway in the micro-simulation model. (Figure 1-1)

Segment: For each VDS, a freeway *segment* refers to the length of freeway from the upstream VDS midpoint to downstream VDS midpoint. (Figure 1-1)

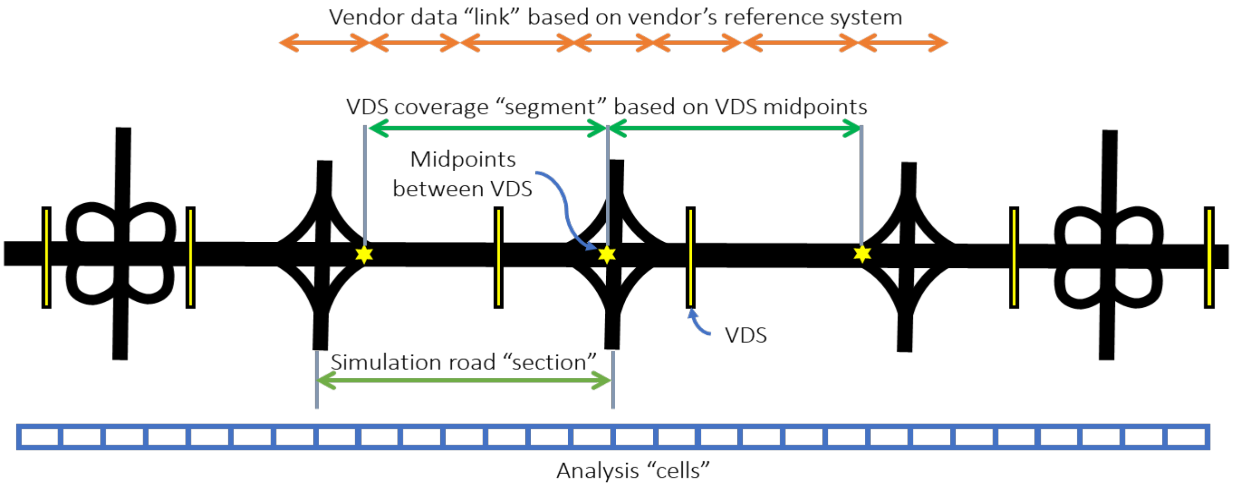

Vendor data "link" based on vendor's reference system
VDS coverage "segment" based on VDS midpoints
Midpoints between VDS
VDS
Simulation road "section"
Analysis "cells"

Figure 1-1 Roadway Nomenclature Diagram

Chapter 1

Introduction

1. INTRODUCTION

This report is the final deliverable for Task Number 3643, Hybrid Data Implementation. This project explores methods to estimate key performance measures in multiple ways using a flexible mix of data, including both traditional sensor data as well as third-party, probe-based mobile data.

Every California Department of Transportation (Caltrans) district generates a quarterly report, called a Mobility Performance Report (MPR)—a report that summarizes key performance measures such as Vehicle Miles Traveled (VMT), Vehicle Hours Traveled (VHT), and Vehicle Hours of Delay (VHD) (Caltrans, 2020b). To compute VHD and VMT, data from Vehicle Detector Stations (VDS) are used. The VDS include 40,000 individual detection zones (Caltrans, 2020a), and maintaining such vast infrastructure requires extensive operational and maintenance support. The availability of third-party, vendor-provided data can augment data from VDS to estimate performance measures, as the required data (such as speed or travel time) can be obtained from third-party vendors.

This report proposes a methodology for using third-party data, investigates advantages and opportunities that come with this data, and provides a roadmap to move forward.

1.1. PURPOSE

The primary purpose of this project is to determine whether and how Caltrans may benefit by incorporating third-party vendor data into its established system for performance measurement. Key goals include the following:

- Reduce costs and increase coverage of traffic monitoring
- Provide a methodology for calculating vehicle hours of delay (VHD)
- Enable smarter deployment of point-based sensors, such as loops
- Provide a roadmap strategy for using third-party data.

1.2. SCOPE

This project considers VMT and VHT estimation but focuses on VHD as the main performance metric of interest. A survey of data offerings by third-party vendors is performed and used to define the characteristics of third-party data.

Algorithms are designed to estimate VHD using a flexible mix of data, including third-party data, point-based sensor data (such as loops), and annual average daily traffic (AADT). The performance of the algorithms is evaluated against a simulated ground truth (SGT) leveraging the Connected Corridors microsimulation model of the I-210 Integrated Corridor Management (ICM) Corridor.

A key challenge is to project (or to conflate) data from multiple sources onto the same domain of analysis to compute performance metrics with high fidelity. Based on the analysis, recommendations are made for improving the meta-data (or configuration data) of Caltrans Performance Measurement System (PeMS).

A framework is presented to determine: (1) improvements possible when fusing third party data to determine delay; and, (2) the error introduced into VHD estimates as point-based sensor data is removed. This framework is then used to evaluate the performance of the algorithms for a range of operating conditions.

Based on the analysis, a strategic roadmap is proposed for incorporating third-party data into PeMS.

1.3. BACKGROUND

Caltrans relies on over 40,000 individual vehicle detection zones to provide information on vehicle data such as flow, occupancy, and speed. This information is in turn used for various system operations and management activities. Gigabytes of data every day is collected and used to provide support for traffic management, real-time traveler information, and system performance monitoring. These functions are vital in supporting Caltrans mission, vision, and goals – Goal 1: Safety and Health, Goal 2: Stewardship and Efficiency, Goal 3: Sustainability, Livability and Economy, and Goal 4: System Performance.

Operating this vast detection system requires extensive resources in the form of engineering and maintenance support along with millions in capital funds to keep them running. Recently, Caltrans programmed over $150 million in State Highway Operation and Protection Program (SHOPP) funds to address failed or failing detection stations across the state.

With the increased availability of third-party probe-based data to provide some of the same data currently obtained through existing detection systems, there should be a renewed effort to look at how those data sources may be able to supplant or augment existing data collection methods. Most third-party data providers can now provide detailed travel time or speed data on any route. In addition, data samples will continue to grow as more cellular devices are used.

To properly integrate these data into the existing reporting platform and into deliverables such as the MPR, research will need to be done to determine how to incorporate the third-party data to provide both real-time and historical performance metrics. This will require evaluating and modifying algorithms currently used in the Caltrans PeMS.

As previously discussed, third-party vendor-provided data can be used by the districts to report performance measures (Bayen, Sharafsaleh and Patire, 2013; Chen and Mei, 2019). For a certain link, the third-party vendors report date, timestamp, link identifier, link length, speed, and travel time (Chen and Mei, 2019). The travel time data is useful for estimating VHD. The feasibility of using third-party travel time data is studied here to evaluate whether it provides any benefit compared to the traditional method where third-party data is unavailable.

Detailed guidelines are needed to outline and describe the data requirements from third-party vendors. These guidelines can help Caltrans develop a roadmap strategy for comparing third-party data vendors and making the appropriate selection. Once the data is procured, data quality monitoring is needed to periodically check the usability of the data. Finally, the framework of VHD estimation needs to be incorporated in PeMS so the districts can use the VHD performance measure whenever needed.

1.4. SUMMARY FINDINGS

All third-party, commercial vendors studied in this project depend on smartphone applications, in-vehicle Original Equipment Manufacturer (OEM) navigational devices, and data from connected vehicles. Traffic speed and travel time data (based on a sample of equipped vehicles) are among the first of the roadway traffic data to become commercially available. In general lane-by-lane, disaggregated speeds are not available. Speeds on High Occupancy Vehicle (HOV) facilities are not reported separately from speeds on mainline freeway lanes. However, as technologies and the use of big data in the transportation industries evolve, new products and services continue to be introduced.

The existing MPR neglects data from on-ramps, off-ramps, and freeway connectors. In other words, performance measures are not calculated on these facilities. As shown in Section 2.4.1, on urban expressways this corresponds to 15%-20% of Caltrans lane-miles.

Any change to the existing methodology for measuring speeds, flows, or travel times will change the estimates for all measures in the MPR. In other words, the integration of third-party data may have a profound effect on all downstream measures that use this data, such as VMT, VHT, bottlenecks, lost productivity, etc.

A key challenge is to project (or to conflate) data from multiple sources, including multiple vendors and PeMS, onto the same domain of analysis to compute performance metrics with high fidelity. PeMS typically provides point-based measurements, such as flows or spot-speeds. Third party data is typically furnished as average speeds or travel times measured over a road link. The coordinate systems used by PeMS and by multiple providers will not, in general, align conveniently.

Using PeMS effectively requires additional meta-data for each VDS, especially at freeway-freeway (fwy-fwy) junctions to specify their locations more precisely. An approach is proposed to organize VDS into groups called Fully Accounted Traffic Volumes (FATV) to support automated validation of sensor data, and to geo-spatially organize sensors for fusion with third-party data.

There are four main methods that were compared for estimating VHD:

- Traditional data and calculation
    - Uses point-sensor data only
    - Calculates delay over long freeway segments
- 3rd party + traditional calculation
    - Only possible when spatial reference systems match
    - Uses point-sensor data for flows and third-party data for travel times
    - Calculates delay over one connector, or a long freeway segment
- Hybrid calculation
    - Required when spatial reference systems do not match

    - Uses point-sensor data for flows and third-party data for travel times
    - Divides long freeway segments into cells for greater accuracy
    - Applies traffic theory to accommodate distance between point-sensors
- Adjustments for limited instrumentation
    - Uses rough estimates for flows and third-party data for travel times

Based on the analysis in this report, the recommended VHD estimation method depends on the infrastructure type and the data available. For freeway mainlines, the best performance was achieved with the hybrid calculation. For HOV lanes, traditional means must be used until third-party data become available that reliably distinguish them from the mainline lanes. For connectors, good performance was obtained using third-party data combined with the traditional calculation. For ramps and arterials, further work is required.

The two main error sources in the traditional data and calculation method are: (1) usage of the g-factor approximation to estimate speeds; and, (2) usage of a point measurement to approximate the measurement across an expanse of road. Single loops, as predominantly deployed, do not actually measure speeds. Even if the point speed is measured with other sensors (e.g., dual loops), the point speed does not reflect the overall operational condition for a freeway segment. For this reason, direct measurement of travel-times possible with third-party data is advantageous.

Based on the analysis conducted in this research, it is found that a hybrid approach provides the best estimates of performance measures. The analysis is conducted for both freeway-mainline and freeway-freeway connectors, and for five different times of day during weekday hours. For almost all the scenarios, the inclusion of travel time data reduces the estimation errors. These benefits hold when the number of fixed detectors is reduced:

- The hybrid method improved VHD estimates when FATV sensor groups were systematically removed (3.4% error with hybrid and 12.7% error with traditional data)
- Third-party data can estimate VHD using AADT on roadways with limited instrumentation, but the error is high (on average 41.3% error when using generic flow profiles, and 19.4% error with measured flow profiles).

In the past, a perceived risk of using hybrid data for traffic management systems was its dependence on external vendors. However, the new risk is that mobile devices are so prolific that drivers are now being influenced by the apps, and traffic management systems lack direct access to this influential, and useful, information.

In an increasingly interconnected world, the future of transportation management will require better and more complete data that can only be obtained through greater connectivity to the data feeds of private vendors as well as increased cooperation and collaboration with local stakeholders. The first step in this direction is to adopt computing tools and infrastructure that have already been tried and implemented at scale in the private sector. Of course, data quality would need to be monitored on a continuous basis, and attention should be paid to costs when selecting a portfolio of data sources. The deployment of a traffic management system that uses hybrid traffic data will provide the opportunity to safely alter Caltrans' strategy for the usage of traditional detectors.

1.5. STRUCTURE OF REPORT

This project was divided into seven project work tasks:

- Task #1: Project management
- Task #2: Background survey
- Task #3: Vehicle Hours of Delay
- Task #4: Point-based deployment strategy
- Task #5: Opportunities for improved coverage
- Task #6: Strategy for incorporation of third-party data
- Task #7: Final Report and Workshop

Task #1 involves project management but no research findings. As such it is not discussed further in this final report.

Task #2 is to provide an overview of the data landscape pertaining to existing data sources in Caltrans' data pipeline, including PeMS (Performance Management System), and third-party, probe-based mobile data, including real-time and historical data needs. The existing data pipeline is discussed in Chapter 2 Section 2. A list of key third-party data vendors and their products are presented in Chapter 2 Section 3, along with a summary of the sources of raw data upon which they depend. In this data review, critical challenges to data fusion and data integrity in the structure of PeMS configuration meta-data are identified and discussed.

The core technical and algorithmic work was conducted in Task #3 and described in Chapter 3. Existing methods for estimating performance measures are described. A review of conflation and data fusion approaches is provided in Chapter 3 Section 2. A framework is described to fuse data from both third-party vendors and Caltrans VDS, the overall method is referred to as the "hybrid method". This framework includes (1) a deep learning-based method to impute missing data in the VDS where data are unavailable, (2) an algorithm to conflate the data from both third-party vendors and VDS along the same freeway, and (3) a method to fuse the two data sources to estimate VMT, VHT, and VHD.

Tasks #4 and #5 involve application of the methods in Task #3 to reveal impacts to the accuracy of performance measures when point-based sensors are removed or replaced by approximate flow data. All of this is presented in Chapter 4.

Task #6 is to formulate a strategy to incorporate third-party data and it is described in Chapter 5. This includes an overall discussion on the advantages and disadvantages of data sources and a proposed organizational concept, called FATV, to augment existing PeMS configuration meta-data.

Finally, this draft final report is the deliverable for Task #7. It is a compilation of the technical memoranda which were reviewed and approved by the Technical Advisory Group for this project. Upon review and approval by Caltrans this Task #7 deliverable will form the basis for discussion in the final workshop.

Chapter 2

Data Landscape

1. OVERVIEW

This chapter provides a background survey for this project, describing the existing Caltrans data pipeline in PeMS, and providing a market review of key third-party data vendors. PeMS typically provides point-based measurements, such as flows or spot-speeds. Third party data is typically furnished as average speeds or travel times measured over a segment of a road. The coordinate systems used by PeMS and by multiple providers will not, in general, align conveniently. Methods are needed to conflate or to project the data onto other coordinate systems for visualization, or for metric calculations.

1.1. SUMMARY FINDINGS

One of the goals of this project is to suggest how performance measurement can be more comprehensive and to provide broader coverage of Caltrans facilities. During the review of the existing methodology, it was revealed that the existing MPR neglects data from on-ramps, off-ramps, and freeway connectors. In other words, performance measures are not calculated on these facilities. On urban expressways this corresponds to 15%-20% of Caltrans lane-miles, as described in Section 2.4.1 of this chapter.

Since many of these freeway connectors and ramps are already instrumented with PeMS VDS, they appear to be easy targets for improving coverage, and achieving these goals of more comprehensive and representative performance measures at a minimal cost. However, delay calculations are typically performed against a speed threshold, and it is unclear what speed threshold might be most appropriate for ramps. On the other hand, ramps are a part of the infrastructure and VMT and VHT are easily interpreted on these facilities.

Any change to the existing methodology for measuring speeds, flows, or travel times will change the estimates for all measures in the MPR. In other words, the integration of third-party data may have a profound effect on all downstream measures that use this data, such as VMT, VHT, bottlenecks, lost productivity, etc. More details on MPR calculation methodology are explained in Section 2.4.2 in this chapter.

Third-party data is typically associated with a map consisting of links and nodes. PeMS data is associated with a linear reference system (not a map). The linear reference system (also known as LRS) has limited expressivity and may create future challenges for data conflation. More detail in Section 2.5.3

All third-party, commercial vendors studied in this project depend on smartphone applications, in-vehicle OEM navigational devices, and data from connected vehicles. Traffic speed and travel time data are among the first of the roadway traffic data to become commercially available. As technologies and the use of big data in the transportation industries evolve, new products and services continue to be introduced. The main vendors and a snapshot of their current offerings are described in Sections 3.2 and 3.3 of this chapter.

2. EXISTING CALTRANS DATA PIPELINE

This section describes the existing sensing and data pipeline used by Caltrans. The focus is on the data pathways that eventually result in the metrics reported in the MPR. Descriptions are simplified to focus on key details relevant to this project.

It is crucial to note that there are two distinct types of data that must be considered:

- Field measurements (data about traffic)
- Meta-data (data about the data)

The field measurements are the flows, speeds, and travel times needed to understand something about traffic conditions. However, the meta-data (sometimes called configuration data) are what is needed to understand how to use the field measurements. The meta-data contains information about where the sensor is located, what side of the freeway it is on, how many lanes there are on the facility, etc. If the meta-data are wrong, then any metrics based on that data are also wrong. Therefore, it is crucial that the meta-data are correct.

If it is intended to fuse traffic data from multiple sources, then the meta-data become even more important. The contents of the meta-data will determine how to project the traffic data onto a common domain of analysis. The meta-data in the Caltrans data pipeline and its linear reference system has limited expressivity, and these limitations are explored further in this report.

2.1. OVERVIEW

Each Caltrans district operates and maintains its own sensing infrastructure. Field elements such as loops and radar connect to a local communications hub. From each hub, data is sent to an Advanced Traffic Management System (ATMS) or central system located in a Transportation Management Center (TMC). In addition, field data are also sent to servers that provide it to PeMS where it is archived and analyzed. Within PeMS, there is an interface from which to generate reports such as the MPR. The MPR is a quarterly report prepared by each Caltrans district that summarizes key performance measures such as VMT (vehicle miles traveled), VHT (vehicle hours traveled), and delay.

2.2. REAL-TIME APPLICATIONS

Fixed sensors embedded in the infrastructure are currently required for traffic operations and control. Typical examples include actuation for traffic signals located at the intersection of freeway ramps and arterial streets, ramp metering control at freeway on-ramps, and real-time pricing for High Occupancy Toll (HOT) lanes.

Emerging real-time applications include ICM projects such as the I-210 Connected Corridors Project. The objective of this pilot project is to:

> …reduce congestion and improve mobility in a section of the I-210 corridor in the San Gabriel Valley of Los Angeles County. This objective will be achieved by coordinating the principal

> elements in the corridor—the I-210 freeway, key surrounding arterials, supporting local transit services, and other relevant transportation systems—and managing them as an integrated and cohesive whole. To attain these operational improvements, the project team will design, develop, implement, and evaluate a pilot Integrated Corridor Management (ICM) system that will help transportation system managers in their decision-making tasks and enable operators, control systems, vehicles, and travelers to work together in a productive and coordinated way. (Connected Corridors, 2020)

At the heart of this ICM is a decision support system that will propose response plans to mitigate traffic congestion resulting from incidents. Response plans may involve reroute guidance, adjustments to ramp meters, and changes to arterial signal control plans to compensate for lost capacity elsewhere in the network.

One requirement for a successful ICM is to collect the data necessary to determine the benefits of the project. Another requirement is having enough situational awareness to know whether a proposed response plan has a good chance of improving traffic. In practice, this means real-time monitoring of traffic conditions (such as flows, densities, and speeds) on mainline, HOV, ramp, and arterial facilities.

2.3. HISTORICAL APPLICATIONS

For the purpose of this report, there are two noteworthy historical applications for PeMS and PeMS-like sensing data:

- Performance measures
- Model building

There are several performance measures and reports generated for Caltrans business purposes, examples include Traffic Census, Monthly Vehicle Miles of Travel, Traffic Volume, and Mobility Performance Reporting. These reports provide key metrics to inform Caltrans decision-making. In addition, ICM projects benefit from some level of modeling. Modeling may be used to facilitate communication with stakeholders, to test out proposed control interventions, or to estimate project benefits.

Model building requires calibration, and calibration requires high quality, self-consistent data. Therefore, PeMS data is a crucial resource for modeling.

2.4. MOBILITY PERFORMANCE REPORT

The MPR basically aggregates data from PeMS and provides it in the form of a report. The key pieces of information that go into the report are measurements from mainline and HOV vehicle detector stations.

Data from other VDS types, such as on-ramps, off-ramps, and freeway connectors are not used. For rural areas, this presents no issue. However, for urban freeways with closely spaced ramps, the MPR leaves out a significant fraction of vehicle miles and vehicle hours on Caltrans right-of-way.

2.4.1. MPR ACCOUNTING OF TRAVEL

A careful accounting of freeway surface area was made possible by leveraging the Connected Corridors I-210 Aimsun model. This model contains detailed geometric information for both freeways and arterial roads contained in the model. Results of the accounting are displayed in Table 2-1.

Table 2-1: Caltrans Facilities in Connected Corridors I-210 Model

	Linear Miles	*Lane Miles*	*Linear Fraction*	*Lane-mile Fraction*
Freeway	48.97	221.75	0.44	0.72
Off-ramp	11.56	17.99	0.10	0.06
On-ramp	12.94	22.77	0.12	0.07
Connector	7.29	13.15	0.07	0.04
HOV	31.04	31.18	0.28	0.10

In terms of linear lane-miles, freeway and HOV facilities make up only 72% of the Caltrans right-of-way in the Pasadena area. The other 28% is made up of ramps and freeway connectors. Even when adjusting for the number of lanes, ramps and connectors still make up 17.6% of Caltrans roads in this area.

The roadway geometry in the Pasadena area is like that in other urban areas. Therefore, the existing methodology neglects approximately 15-20% of the lane-miles in urban areas throughout the state of California. In other words, a significant fraction of VHT and VMT on Caltrans ROW is not accounted for. For mostly rural districts, the error is probably small. However, in districts with large urban areas the error could be substantial.

2.4.2. MOBILITY PERFORMANCE REPORT CALCULATIONS

In terms of configuration meta-data, each VDS is responsible for representing a specified length, of freeway. In terms of traffic data, the two key measurements are five-minute counts and speeds. These three values (length, count, and speed) are used to generate all of the measures in the MPR (Caltrans, 2012).

VMT is represented in the units of [vehicle·miles] and can be calculated over an arbitrary time interval. Delay is represented in units of [vehicle·hours] and is calculated against a threshold speed. Mathematical expressions are provided in Section 3.1.2 of Chapter 3, Estimating Vehicle Hours of Delay.

Delay is a key metric that is then used to calculate the cost of congestion, lost time, and wasted fuel. The lost lane miles depend on measuring the count of vehicles at a cross-section of a freeway. Bottleneck locations are detected algorithmically with a speed threshold and a VHD threshold.

Therefore, any change to the existing methodology for measuring speeds or flows will change the resulting estimates for all measures in the MPR.

A key point to make is that if the goal is to save money by reducing the number of point-detectors and supplementing where possible with third-party data, it is necessary to add information to the inventory of assets. As explained in the next section, the crucial information that will be needed is a network map of Caltrans sensing facilities, and appropriately updated configuration meta-data.

2.5. PEMS

PeMS (Performance Measurement System) is an archive of California freeway data maintained by Caltrans headquarters. While each individual district has its own systems and sensors, each sensor provides real-time data to centralized servers that eventually end up in PeMS.

2.5.1. LINEAR REFERENCE SYSTEM

Caltrans has a well-established linear reference system for the purpose of describing locations on the freeway. This system makes use of so-called CA PM (California Postmiles) and Abs PM (Absolute Postmiles). The CA PM are best conceptualized as labels (not monotonic post miles) with letters and numbers to indicate a place name. The Abs PM are monotonically increasing post miles that span the length of the freeway.

2.5.2. SENSOR ORGANIZATION META-DATA

The logical organization of PeMS data (and meta-data) follows directly from the physical organization of the field equipment. Control boxes containing power and communications equipment are installed along the freeway and are assigned to linear postmile locations. These locations are also mapped to a pair of latitude and longitude points snapped to the right-of-way centerline (typically in the median between two directions of the freeway).

The control box may support any number of PeMS sensors. Sensors are predominantly loop detectors, but other sensor types such as radar are also used to collect data. Loops, for example. might be placed on mainline lanes, HOV lanes, off-ramps, on-ramps, etc. However, they will inherit their position (postmile, latitude, and longitude) from their control box. In other words, the true, real-life, physical geo-location (latitude and longitude) of the sensor is not the same as its reported position (latitude and longitude) provided in the PeMS meta-data. This is a crucial point, and it is also a limitation in the expressivity of the meta-data.

The meta-data takes the form of a spreadsheet with the following columns: Fwy, District, County, City, CA PM, Abs PM, Length, ID, Name, Lanes, Type, Sensor Type, HOV, MS ID, and IRM. The typical usage for the "Fwy" column is to indicate the physical freeway where the sensor is located. The "Name" column is used to indicate the closest parallel street or feature. The "Type" column is used to indicate the relationship of the sensor to the physical freeway in the "Fwy" column. This convention is adequate for typical rural freeways with few major fwy-fwy interchanges. However, this convention fails at major fwy-fwy interchanges common to urban freeways.

As described below, the data fields are used differently at interchanges. A “correct” entry for a data field may not exist given system limitations. As a result of inconsistent usage patterns, it is:

1. Difficult to check data consistency and data quality;
2. Difficult to use the data for modeling in ICM applications; and,
3. Difficult to use the data when incorporating third-party data.

2.5.3. CHALLENGES AT FREEWAY INTERCHANGES

The linear reference system is reasonable for long, rural freeways with simple geometries. However, it becomes difficult in situations involving complicated freeway junctions. Figure 2-1 shows the location of VDS sensors at the I-210 & I-605 junction in Los Angeles County. Notice that one blue place-mark coded as a mainline sensor is located on the fwy-fwy connection between I-605 NB and I-210 WB.

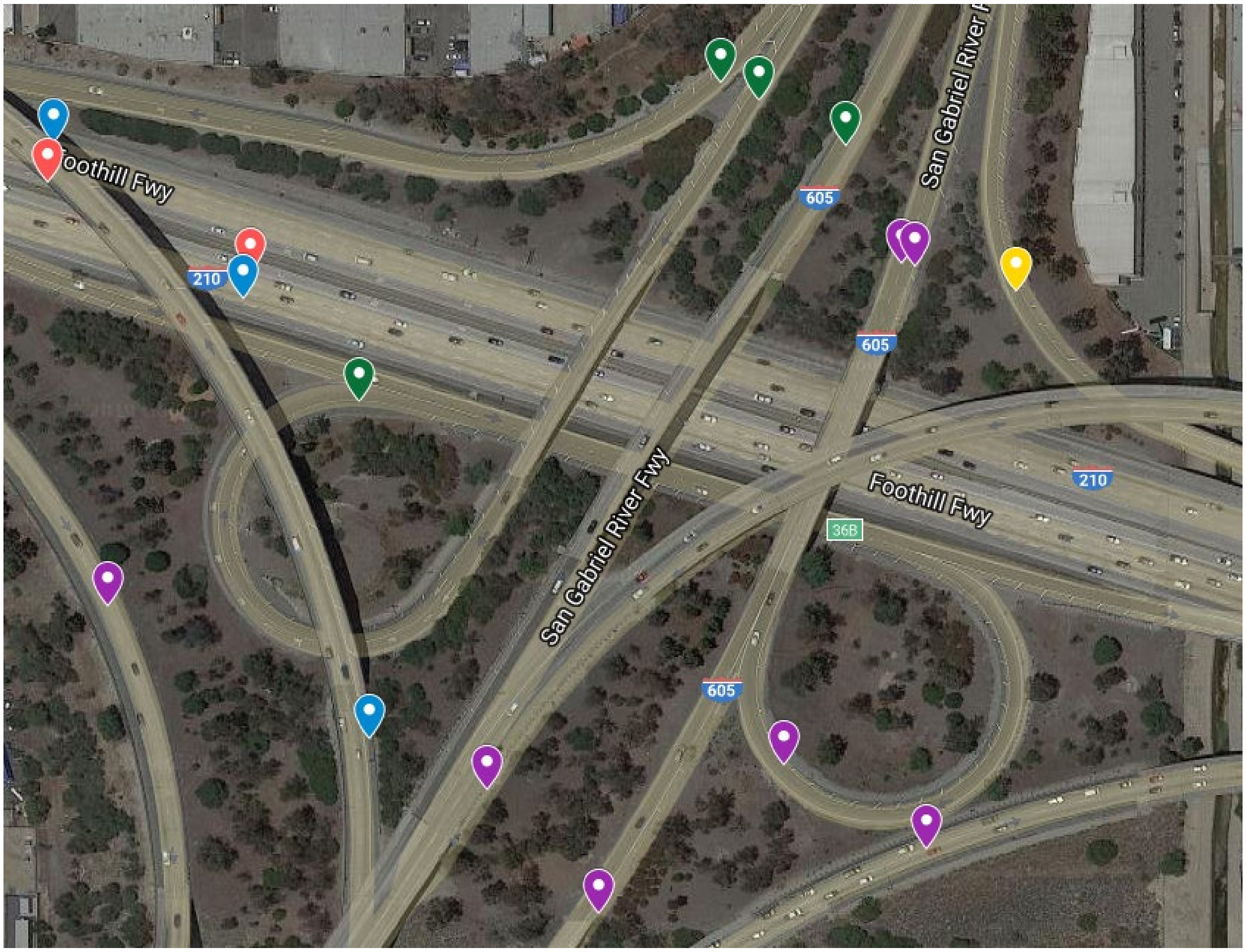


Figure 2-1: Junction between I-210 and I-605 illustrating physical locations of VDS sensors. Place-mark colors blue, red, yellow, green, and purple indicate VDS types mainline, HOV, off-ramp, on-ramp, and fwy-fwy connector, respectively.

PeMS provides a schematic diagram called a strip-map to serve as a visual aid to locate VDS. Figure 2-2 shows the strip-map for I-210 in the vicinity of the junction at I-605. Notice that the WB direction on the top of the strip-map appears to constrict to two lanes, whereas the satellite view in

Figure 2-1 shows four lanes plus one HOV lane all the way through.

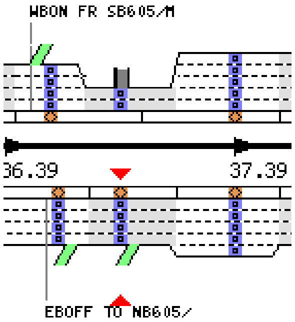


Figure 2-2: Strip-map of I-210. WB direction is on top and EB direction is on the bottom. Numbers in the center indicate Abs PM. VDS are shown as blue, pink, green, or grey markings.

PeMS configuration meta-data in this area of I-210 are displayed in Table 2-2. The list of all VDS associated with controller MS ID 2407 located at Abs PM 36.89 on westbound I-210 is included. Inspection of the "Name" column in Table 2-2, and cross-referencing with Figure 2-3, reveals that most of these VDS have nothing to do with pavement located on westbound I-210. Two possible exceptions might be VDS 773206 and 775795. The former measures SB 605 flows of which some may have originated from WB 210. The latter measures flow on the connector between NB 605 and WB 210. The categorization of VDS 775795 may be responsible for the strip-map representation showing the I-210 WB freeway having only two through lanes.

Table 2-2: Listing of VDS associated with MS ID 2407

Fwy	*County*	*CA PM*	*Abs PM*	*ID*	*Name*	*Lanes*	*Type*	*MS ID*
I210-W	Los Angeles	R36.6	36.89	773204	NB 605 TO MT. OLIVE	1	Fwy-Fwy	2407
I210-W	Los Angeles	R36.6	36.89	773205	EB 210 TO MT. OLIVE	1	Fwy-Fwy	2407
I210-W	Los Angeles	R36.6	36.89	773206	SB 605 FROM WB 210	2	Fwy-Fwy	2407
I210-W	Los Angeles	R36.6	36.89	773207	NB 605 TO EB 210	2	Fwy-Fwy	2407
I210-W	Los Angeles	R36.6	36.89	775795	NB 605 TO WB 210	2	Mainline	2407
I210-W	Los Angeles	R36.61	36.90	775796	EB 210 TO SB 605	2	Fwy-Fwy	2407

The actual physical locations of the VDS in Table 2-2, above, are shown in Figure 2-3, below. The six VDS overlap on the strip-map so that only two (775795 and 773204) are visible in Figure 2-2.

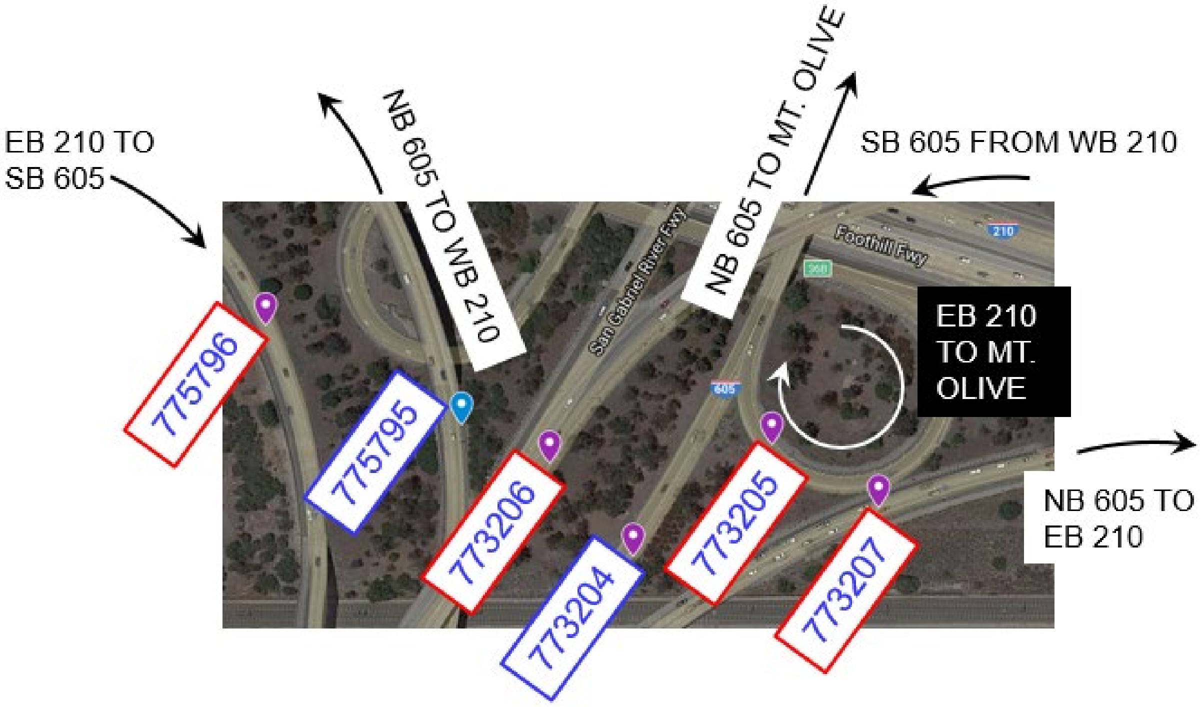


Figure 2-3: Physical locations of VDS from Table 2-2. Of these six VDS, only VDS 775795 and 773204, outlined in blue, appear in the mouse-over function on the online version of the strip-map.

Figure 2-4 shows the strip-map for I-605 in the vicinity of the junction at I-210. Notice that the SB direction on the bottom of the strip-map appears to have VDS at an offramp, at a fwy-fwy connector, and at an on-ramp at about the same place, whereas the NB direction has no VDS.

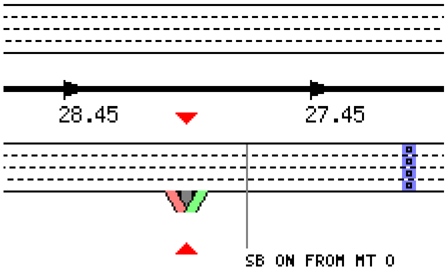


Figure 2-4: Strip-map of I-605. NB direction is on top and SB direction is on the bottom. Numbers in the center indicate Abs PM. VDS are shown as pink, green, or grey markings.

Table 2-3 displays PeMS configuration meta-data for the VDS associated with controller MS ID 4430 located at Abs PM 27.95 on southbound I-605. Inspection of the "Name" column in Table 2-3, and cross-referencing with Figure 2-5, reveals that most of these VDS have nothing to do with pavement located on southbound I-605. The main exception is VDS 774260, which is categorized reasonably.

Table 2-3: Listing of VDS associated with MS ID 4430

Fwy	*County*	*CA PM*	*Abs PM*	*ID*	*Name*	*Lanes*	*Type*	*MS ID*
I605-S	Los Angeles	25.9	27.95	774264	NB 605 TO MT OLIVE	1	Fwy-Fwy	4430
I605-S	Los Angeles	25.9	27.95	774262	WB 210 TO MT OLIVE	1	Off Ramp	4430
I605-S	Los Angeles	25.9	27.95	774263	MT OLIVE TO WB 210	1	On Ramp	4430
I605-S	Los Angeles	25.9	27.95	774261	EB 210 TO MT OLIVE	1	Fwy-Fwy	4430
I605-S	Los Angeles	25.9	27.95	774260	MT OLIVE TO SB 605	1	On Ramp	4430
I605-S	Los Angeles	25.9	27.95	774258	MT OLIVE TO EB 210	1	On Ramp	4430

The actual physical locations of the VDS in Table 2-3, above, are shown in Figure 2-5, below. The six VDS overlap on the strip-map so that only three (774262, 774261, and 774258) are visible in Figure 2-4.

When comparing Table 2-3 to Figure 2-5 it is crucial to note that the "Type" provided in the table does not, in general, describe the relationship of the sensor to the "Fwy" as it would for typical installations away from a major fwy-fwy interchange. For example, while VDS 774262 could reasonably be described as an off-ramp with respect to WB I-210, it is not an off-ramp with respect to SB I-605. Unfortunately, the PeMS configuration specifies VDS 774262 as an off-ramp with respect to the wrong freeway as shown in Table 2-3. As a result, this meta-data is difficult to use.

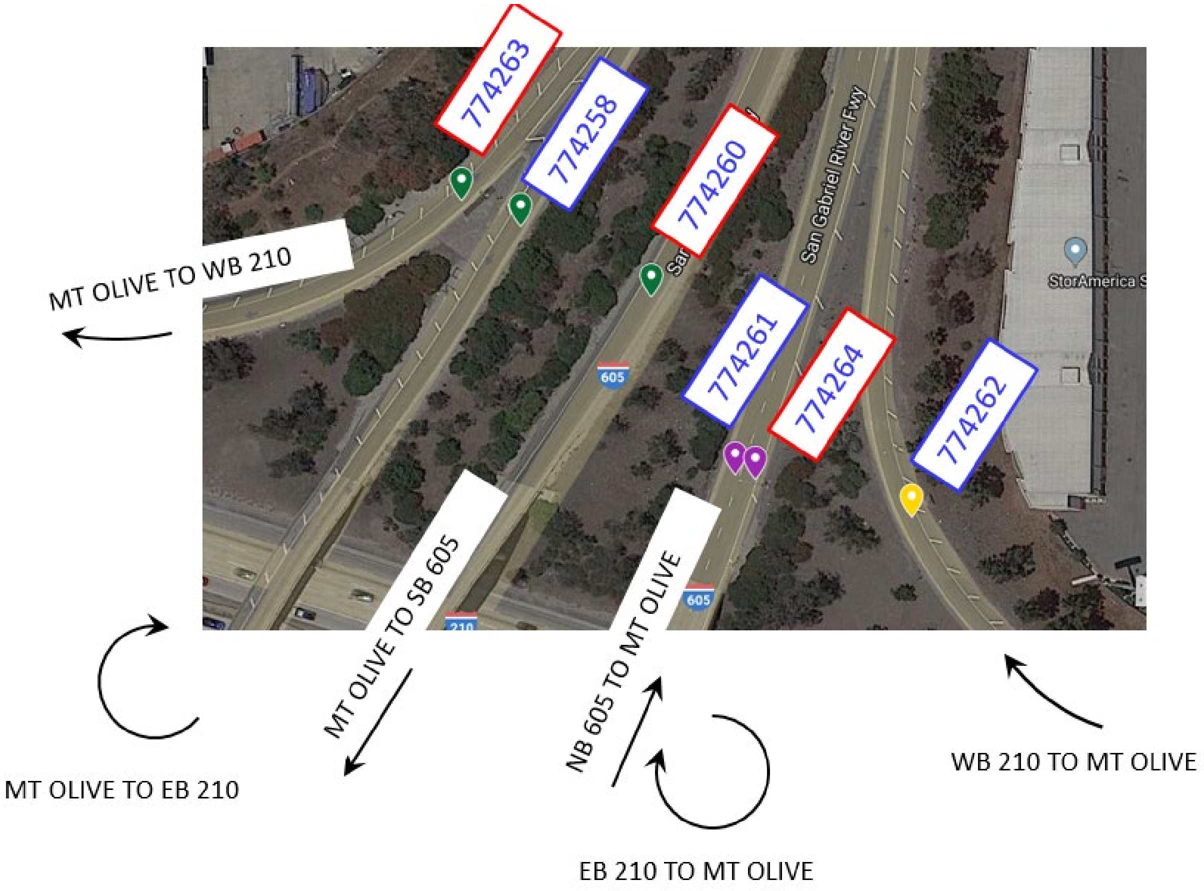


Figure 2-5: Physical locations of VDS from Table 2-3. Of these six VDS, only VDS 774262, 774261, and 774258, outlined in blue, appear in the mouse-over function on the online version of the strip-map.

Subsequent sections will investigate how third-party data may be used to expand coverage and improve the quality of Caltrans' performance reporting capabilities. The first natural extension is to expand coverage to ramps and connectors that already have PeMS detectors but are not yet represented in the performance measures. The first step toward this would be to improve the configuration meta-data of PeMS VDS on ramps and freeway connectors so that they can be appropriately conflated/mapped with third-party data available on the same facilities.

3. SURVEY OF THIRD-PARTY DATA

3.1. TAXONOMY OF DATA COLLECTION METHODS

Traffic data collection methods can be organized into three categories, each with its own strengths and limitations. (Figure 2-6)

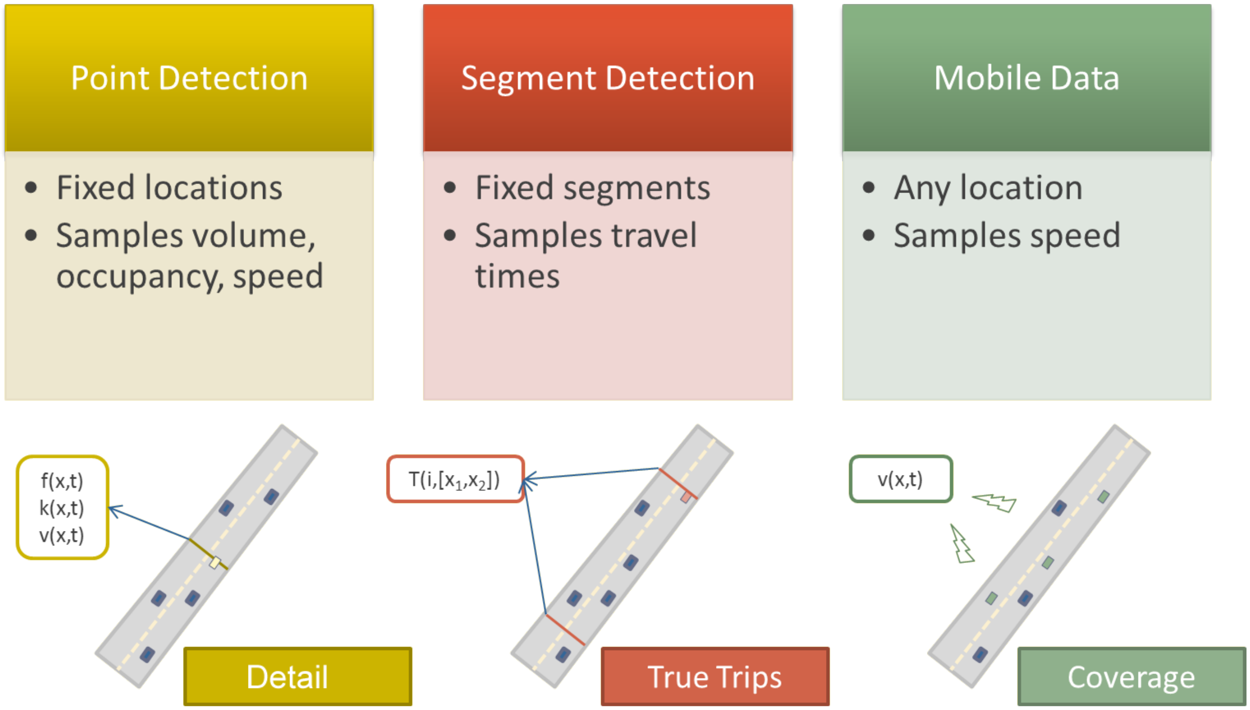


Figure 2-6: A comparison of traffic data collection methods (Bayen, Sharafsaleh and Patire, 2013)

3.1.1. POINT-BASED COLLECTION METHODS

Point-based data collection methods, such as inductive loops and radars, measure traffic flows and/or speeds at one dedicated location. The strength of these methods is that they capture the complete cross-section of all vehicles passing by a given location, and therefore obtain reliable measures of flow and speed—within the capabilities of each technology. The disadvantage is that they provide no direct information about what happens between those locations. For example, there is no way to detect a traffic incident between two point-detectors until a change in traffic state (resulting from the incident) propagates upstream or downstream to the point detectors. Even then, the exact position of the incident somewhere between the two point-detectors would remain unknown.

3.1.2. SEGMENT-BASED COLLECTION METHODS

The second category of traffic data collection methods provides trip times for preset road segments. Segment-based data collection is achieved by vehicle re-identification, that is, the ability to uniquely match records of a traveling vehicle obtained at two different locations. Examples in this category include:

- Toll-tag readers
- License plate readers
- Magnetometers
- Bluetooth MAC address readers
- WIFI MAC address readers

Each of these methods can measure travel times between two locations. The number of vehicles that get re-identified varies with the specifics of each technology, and its deployment. With Bluetooth and WIFI, this rate also depends on the prevailing penetration rates of these technologies in consumer devices. In practice, the sample size is enough to provide useful median travel times. As with the point-based collection methods, segment-based methods require dedicated field infrastructure.

3.1.3. MOBILE DATA SOURCES

The third category of traffic data collection methods relies on the proliferation of GPS-enabled mobile devices and data networks to extract the position of individual vehicles over time. This offers two key advantages: (1) no field equipment is necessary (save for cellular network infrastructure, but that is exogenous), and (2) data can be obtained from virtually any location on the roadway network where cellular coverage exists. Mobile data sources can be further divided into several categories:

i. **Smartphone applications:** GPS-enabled smartphones running any location-based app provide their location information. Depending on the app, the rate of location updates may vary. This app-based data collection method is one of the main data streams of INRIX.

ii. **In-vehicle navigation devices:** GPS-enabled devices embedded in the vehicle's dashboard provide predictive navigational aid (e.g., visualizations of the vehicle's current location and route-finding services). The services offered by the in-vehicle navigational devices are very similar, if not identical, to the smartphone applications. Some of the in-vehicle navigational devices also include safety features, such as the ability to call "911" automatically if the vehicle is involved in a collision. The major distinction between the in-vehicle devices and smartphone applications is that the in-vehicle navigational devices are marketed, sold, and installed in-vehicle by the vehicle manufacturers.

iii. **Fleet telematics:** Operators of vehicle fleets (including commercial trucking operations, rental car fleets, taxi fleets, transit bus fleets, etc.) track each vehicle's position (provided by GPS units installed in the vehicle). Many of these fleets agree to let traffic information aggregators use that data to estimate current traffic conditions and archive it for historical reference.

iv. **Connected Vehicles**: High-end, luxury vehicles have telematics modules that collect data about the vehicle and its internal diagnostics. Connected vehicles are vehicles that use any of a number of different communication technologies to communicate with the driver, other cars on the road (vehicle-to-vehicle [V2V]), roadside infrastructure (vehicle-to-infrastructure [V2I]), and the "Cloud" [V2C] (Center for Advanced Automotive Technology, no date). The shared or transmitted data can include the vehicle's GPS position, speed and heading, acceleration and braking data, the vehicle identification and type of vehicle, along with information about the vehicle's current operating conditions from the vehicle's internal diagnostics systems.

3.2. REVIEW OF KEY VENDORS

This section provides a brief introduction to key data sources and/or vendors of mobile traffic data and summarizes their commercially provided data products and services.

All of the described commercial vendors listed below depend heavily on mobile data sources (smartphone applications, in-vehicle navigation devices, and connected vehicles) as their main data sources for providing empirically-based traffic speeds and travel time information along with the other commercially available roadway performance measures. Many of the vendors depend on secondary data sources (like State DOT provided traffic volume and speed data from roadway sensors, weather and/or incident data) to enhance services provided and/or for validation purposes. Even though these "big-data" vendors are vying for market share, it is not uncommon to see data sharing and collaborating or data sharing agreements between subsets of these vendors.

3.2.1. FHWA NPMRDS

The National Performance Measure Research Data Set (NPMRDS) is a dataset acquired by FHWA for use in transportation performance measurement. The NPMRDS contains empirically based traffic passenger (auto) and commercial freight speeds and travel times on a set of predetermined roadway segments that are part of the U.S. National Highway System and for 25+ key Canadian and Mexican border crossings. The NPMRDS is the default dataset for calculating the new US Federal 'PM3' system and freight performance measures.

The first NPMRDS dataset was provided by HERE North America, LLC (formerly known as Nokia/NAVTEQ). Starting in 2017, the NPMRDS data have been provided by INRIX. The NPMRDS is made available free of charge to State Departments of Transportation and Metropolitan Planning Organizations to use for their performance management activities.

3.2.2. HERE TECHNOLOGIES (HERE, 2019)

HERE Technologies was founded as Navteq in Sunnyvale, California in 1985, and provides mapping and location data and related services to individuals and companies. In 2007, the company was acquired by Finland-based Nokia. Currently HERE is headquartered in Amsterdam, Netherlands, and is majority-owned by a consortium of German automotive companies. HERE offer clients a range of products including:

- Automotive Products – auto/mobile SDK for connected embedded navigation solutions, real-time navigational data and services, anticipatory data and sensor support for ADAS and autonomous driving applications, weather data, locational EV charging station data, locational fuel price data, locational parking availability data, hazard warnings, intelligent sensor data for autonomous driving solutions, real-time traffic data

- Location Services Products – fleet telematics, geocoding (mapping of geo-coordinates and addresses), interactive geo-visualization services, mobile SDK, interactive mapping, places and routing data, services, and products.
- Map Content and Positioning Products – map data with visual places footprints, driver maneuver assistance (in-vehicle guidance for upcoming exits and lane splits) and smart positioning for mobile devices.
- Traffic Products – real-time and historical traffic data, traffic analytics, and dashboards.

3.2.3. TOMTOM (TOMTOM, 2019)

TomTom was founded in 1991 and is headquartered in De Ruijterkade, Amsterdam, Netherlands. TomTom has offices in 30 countries. TomTom is a Dutch multinational developer and creator of location technologies and consumer electronics. Since 2008, TomTom has been collecting anonymous consumer-driven GPS based measurements worldwide and used to build its historical traffic database.

TomTom's products include applications and products to aid drivers (navigation device and trip apps and devices), the automotive industry (autonomous driving apps and support, HD maps and map data for autonomous and traditional vehicles), and fleet management (enabling fleet management, vehicle tracking, fleet optimization, workforce management, green and safe driving, and business integration) business solutions and products.

3.2.4. INRIX (INRIX, 2019)

Founded in 2005, and headquartered in Kirkland Washington, INRIX has about 350 employees. INRIX collects anonymized data on traffic congestion, traffic incidents, parking, and weather-related road conditions from millions of data points daily in over 80 countries. These data are combined and aggregated from in-vehicle devices and mobile devices, Departments of Transportation traffic data, cameras and sensors on roadways, and major events expected to impact traffic.

INRIX provides a variety of products, apps, and solutions for drivers, the automotive and trucking industries, and government agencies and their business partners, including:

- INRIX Drive Time – provides real-time assessment of potential commute and travel times.
- Roadway Analytics – Data as a service platform and tools to optimize roadway planning, performance monitoring, and the decision-making process.
- Performance Measures – Transportation data and intelligence for public agencies to help optimize roadway planning and decision-making.
- Population Analytics – combines both GPS and mobile data intelligence to analyze and provide an understanding of the movement of people.
- Volume – a traffic volume dataset with (U.S.) nationwide coverage across 2.65 million miles of road that includes vehicle volume by street direction, time of day, and day of week (in 15-minute bins by road segment).

- Trips – Origin-destination data to better understand the movement of people and the trips they make.

3.2.5. STREETLIGHT (STREETLIGHT DATA INC., 2019)

StreetLight Data was founded in 2011 and is headquartered in San Francisco, California. Every month, Streetlight Data takes in, indexes and processes over 100 billion anonymized location records from smartphone apps and in-vehicle navigation devices, and additional data from numerous other sources like parcel data and digital road network data. StreetLight Data validates the resulting traffic speeds, volumes and travel patterns against Department of Transportation traffic counters and embedded sensors data. Additionally, StreetLight Data fuses and enriches their datasets using supplemental data, like transit ticketing, shared mobility, or IoT data. Lastly, StreetLight Data normalizes and aggregates the data into analytics, delivering empirically based data products on the movements of vehicles, bicycles and pedestrians.

StreetLight Data's traffic-related data products offered to private and public agency clients, include: trip speed, duration, and length, trip purpose, origin-destination metrics, and AADT (counts); travel modes include: bicycle, pedestrian, personal vehicles, ride-hailing and delivery, and truck trips.

3.2.6. CITILABS (CITILABS, 2019)

Headquartered in Sacramento, California with offices in Atlanta, Tallahassee, Abu Dhabi, and Milan, Citilabs provides a comprehensive suite of transportation industry related products, services and solutions to public and private clients. Citilabs supports nearly 2,500 clients in more than 70 countries.

Citilabs has long been associated with travel demand modeling software, services and solutions. Nonetheless, in recent years Citilabs has expanded their set of products to include big-data transportation data and analytics for private and government agency clients. With their Streetlytics platform, Citilabs can provide empirically based (historical) traffic volumes and speeds on nearly all public roadways in California. The Citilabs' Streetlytics platform pulls data and information from billions of points of GPS, cellular, connected car, Bluetooth, ticketing, demographics, and ground truth data to produce traffic-related utilization and performance metrics on public roadways. To accomplish this, Streetlytics employs a proprietary optimization process, which combines data of multiple types from multiple sources:

- Sampled location data from the movement of smartphones and vehicles
- Full population movement data calculated using models of travel behavior applied on current household and employment data
- Ground truth measurement from a database of current traffic counts

The Citilabs Streetlytics platform's key features and services include:

- Directional speed and volume data for roadways (including minor arterials and collector streets); hour-by-hour data by weekday type and month of year
- Trip purpose and mode of travel data
- Route or itinerary data – routes used to travel between origins and destinations

- Home location and demographic characteristics of travelers

3.3. COMPARISON OF PRODUCTS

This section presents a matrix comparing key products and data delivery capabilities of each of the commercial vendors.

All the commercial vendors of “big-data” roadway traffic utilization and performance data depend heavily on smartphone applications, in-vehicle OEM navigational devices, and data from connected vehicles. To obtain these data, the vendors have business agreements with multiple cell phone manufacturers, carriers, and/or smartphone app providers. For example, INRIX has a free downloadable app aptly named “Inrix Traffic” which provides maps, navigational or route guidance information, and driver alerts. HERE Technologies is majority-owned by a consortium of German automotive companies – as such, they have unique data sharing opportunities with these auto manufacturers.

The matrix in Table 3-1 summarizes the vendor’s data sources and relevant data products, along with a few supplemental information categories of interest. In Table 3-1, “CELL/GPS/CV” indicates that the data sources were from the suite of smartphone applications, in-vehicle OEM navigation devices, and from connected vehicles. Note that HERE provides “split lane speeds” which are speed estimates for two dissimilar lane groups, where the speed on one lane group differs from the speed on the second lane group. Typical locations are freeway diverges and freeway merges where one set of lanes or lane group is congested, and the adjacent lane group may be freely flowing.

The viability of each of the vendor’s products depends on several factors, like:

- the importance of real-time data vs only requiring historic data
- whether volume data are required, or if speed (and vehicular travel time data) will suffice
- whether the data are required for interstate and freeways only (i.e., where Caltrans PeMS data are available) or whether the data are required for arterial and/or less traveled roadways
- the required accuracies of the data to meet Caltrans’ needs

Aggregated traffic speed and travel time data were among the first of the roadway traffic data that became commercially available. The ability to provide trip origin-destination estimates came several years later. Providing traffic volume estimates is a relatively new feature for these “big-data” providers, only available within the last few years. As the popularity of cell phones and vehicle route guidance apps grew, the amount of data available to these vendors increased, as did the reliability and accuracy of their traffic speed and travel time estimations improved (and the listing of products offered expanded). As technologies and the use of big data in the transportation industries continue to evolve and advance, costs will decline and products will become even more reliable, robust, and comprehensive.

Table 3-1: Comparative Summary of Traffic Data Provider

	FHWA NPMRDS	HERE Technologies	TOMTOM	INRIX	STREETLIGHT	CITILABS
Key Data Sources	HERE Technologies or INRIX	CELL/GPS/CV Multiple sources	CELL/GPS/CV Multiple sources	CELL/GPS/CV Multiple sources	CELL/GPS/CV Multiple sources	CELL/GPS/CV (AirSage) Agency Traffic Counts
Data Collection Method(s)	HERE Technologies or INRIX Methods	Smartphone App, Vehicle OEM device, and Multiple Other	Smartphone App, Vehicle OEM device, and Multiple Other	Smartphone App, Vehicle OEM device, and Multiple Other	Smartphone App, Vehicle OEM device, and Multiple Other	AirSage and Citilabs proprietary optimization process
Main Products	(Auto and Truck) SPEED	SPEED	SPEED	SPEED VOLUME	SPEED VOLUME O-D	SPEED VOLUME O-D
Additional Data Products and/or Information	Historical speeds and travel times (auto and truck modes)	Historical speeds and travel times, Real-time speeds and travel times, HOV lane speeds, split lane speeds, incident feed, traffic safety warnings	Historical speeds and travel times, Routes, O-D, Incidents, Bottlenecks	Historical speeds and travel times, Real-time speeds and travel times, Bottlenecks, O-D, Volume, Parking, Population Warehousing	Historical speeds and travel times, Trip Duration, Trip Length, Trip Purpose Vehicle AADT, O-D, (Ped/Bike estimates)	"Streetlytics" historical speed And travel times, O-D by block-group, AADT & Hourly Traffic Volumes (Trip Purpose and Mode estimates)
Real-time Delivery Capability	NO	YES Real-time and predictive	YES Real-time and predictive	YES Real-time and predictive	NO	NO
Historical Delivery Capability	YES (delivered monthly)	YES (delivered daily)	YES	YES (delivered daily)	YES (delivered daily)	YES (as per client agreement)
Data validation reports?	YES	YES	?	YES	YES	YES
Mapping Capability	NO (uses HERE mapping)	Have map products	Have map products	Previously used OSM (free) and TomTom (premium), migrating to HERE	NO (uses INRIX mapping)	NO (uses HERE mapping)

Chapter 3

Estimating Vehicle Hours of Delay

1. OVERVIEW

This chapter focuses on developing a reporting method for Vehicle Hours of Delay (VHD) and algorithms to estimate it using a flexible mix of probe data from third-party vendors and data from traditional fixed detectors. Other performance measures, such as Vehicle Miles Traveled (VMT) and Vehicle Hours Traveled (VHT) are also considered.

1.1. APPROACH

The overall framework incorporates four steps as shown in Figure 3-1. At first, the data is acquired from both traditional point detectors or VDS, and third-party vendors. An initial data quality check is conducted to evaluate whether the data are usable to estimate performance measures. After performing the quality control on the data, imputation may be performed if data are missing. Both flow and travel time data are conflated to project them onto the desired cell. After having both flow and travel time data conflated, data fusion is performed to calculate the desired performance measures.

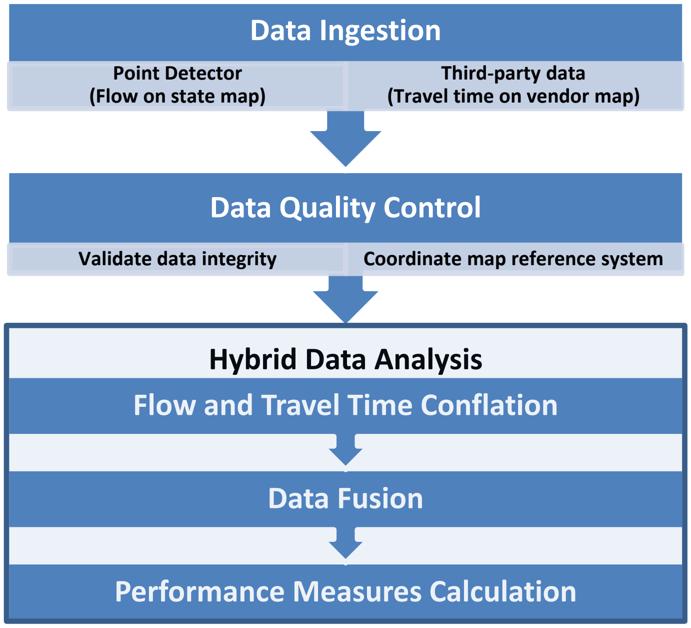


Figure 3-1 Steps for performance measurement estimation

Figure 3-2 presents the data conflation and fusion steps with a simple schematic diagram. It shows that data from PeMS VDS (Figure 3-2a) are available on specific points along the freeway, which does not, in general, line up with the layout of the third-party vendor provided data (Figure 3-2b). Both VDS and third-party data can include imputed data, in case the real-time measured data is unavailable due to detector malfunction, communication error, or absence of probe vehicles. Often the imputed data are drawn from historic observations over the same spatio-temporal domain. Once data are conflated on the same network (Figure 3-2c), both flow and travel time data are available for each cell. Later using the conflated data, performance measures are estimated on the cells, and aggregated over the total spatio-temporal coverage of interest (Figure 3-2d).

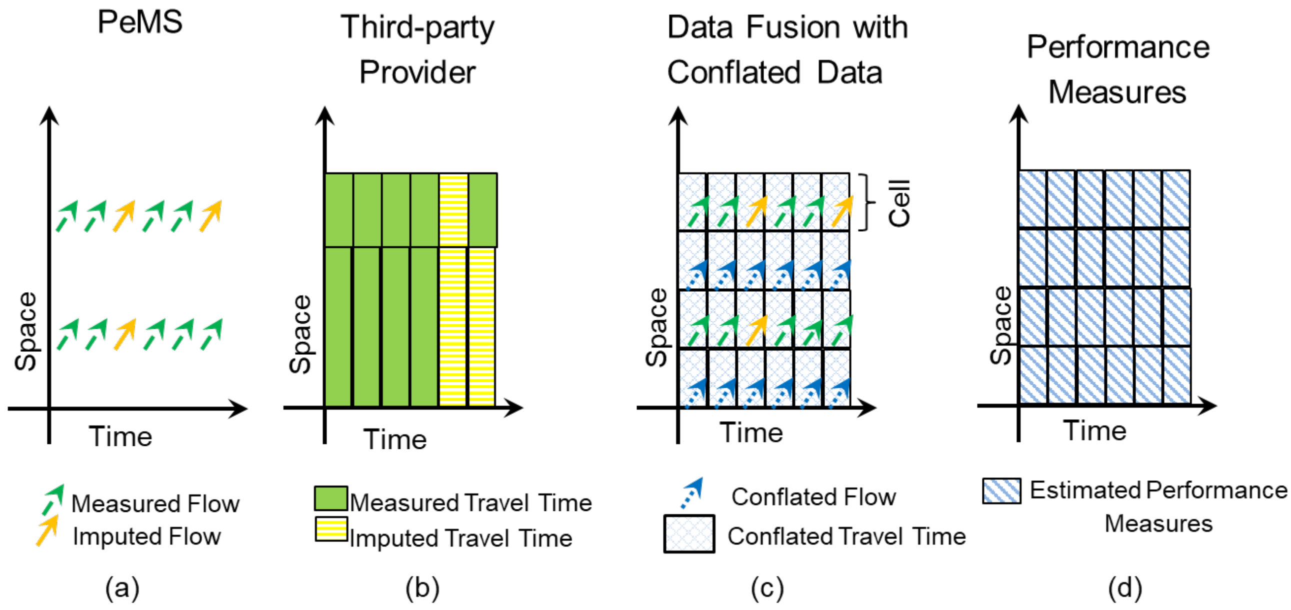


Figure 3-2 Data conflation and fusion

In this report, the evaluations of the developed algorithms are performed using a simulation model. Using any simulated network, the overall framework (as shown in Figure 3-3) can be implemented. A simulated model generates data based on the data characteristics of PeMS and any representative third-party vendor, (e.g., vendor A and B as shown in the figure). The intermediate steps of data quality control, imputation, conflation, and fusion follow the sequence of Figure 3-1. In this research, a single vendor is considered, and imputation is not included in the evaluation of final performance measures. More details on the conflation and fusion criteria are included in Section 3.

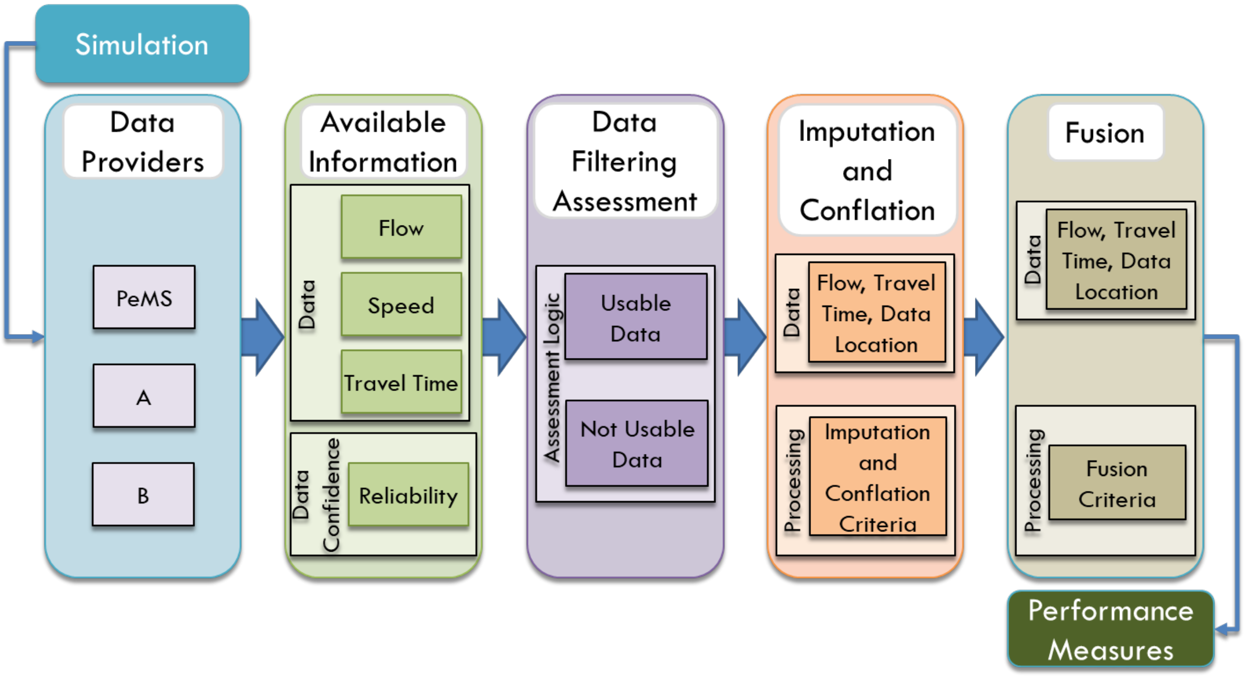


Figure 3-3 Research approach

1.2. SUMMARY FINDINGS

This section discusses a framework of estimating performance measures using a mix of data from traditional point detectors and third-party vendors. For the freeway mainline analysis, probe penetration is 5%, whereas for connector analysis all vehicle data are used. For both analyses, point detector speed is estimated with a g-factor based method, which is discussed in Section 3.1.1.

Based on the analysis conducted in the research, it is found that the hybrid method provides a better estimate of performance measures compared to the single point detector-based method. The analysis is conducted for both freeway-mainline and freeway-freeway connectors, and for five different times of day during weekday hours. For almost all the scenarios, similar findings are observed that the inclusion of third-party vendor-provided data reduces the estimation errors.

2. LITERATURE REVIEW

This section provides a review of related research involving methods for conflation and data fusion.

2.1. DATA CONFLATION

In one study, the authors presented a smoothing method which can interpolate data from single stationary sensors to any intermediate points on a spatio-temporal domain (Treiber and Helbing, 2002). The authors reconstructed data from incomplete information to identify bottleneck efficiently. They used a non-linear weight-based reconstruction method, which used both congested and free-flow velocity information as the a priori traffic estimate. They derived two linear anisotropic kernel functions to smooth the available data based on traffic speed propagation in two states (i.e., free-flow and congested). Fixed values were used for spatio-temporal smoothing window, perturbation propagation velocity, and transition velocity (for both free-flow and congested). The method was used to identify bottlenecks in two real-world scenarios. The authors reconstructed speed for a freeway section using only 35% of the information. However, only visual comparison was used to evaluate the reconstructed data. They suggested the incorporation of traffic parameter-continuity equations to increase reconstruction accuracy.

In another study, the authors extended the method developed by Treiber & Helbing (2002) to provide a heuristic data fusion model that follows traffic flow theory, and can reconstruct the data with inherent structural ambiguity in the spatio-temporal domain (Van Lint and Hoogendoorn, 2010). As first order traffic flow models and Kalman Filters can only be used to fuse data that are already aligned, the authors developed the Extended Generalized Treiber Helbing filter. They applied an area-based restriction to reconstruct data at any specific point. To fuse data from multiple data sources, they also developed a linear formulation with one additional weight. This weight considered the reliability of the data sources. A simulation study was used with 19 km freeway segment in Europe with on-ramp, off-ramp and weaving sections. Compared to the single data source-based reconstruction, bias was reduced while using data from multiple sources. It was found that the dense loop placement (inter-detector distance of 500m) with (1) floating car data (2%, 5%, and 10%) or (2) automatic vehicle identifier (AVI) (coverage: 1500m and 3000 m) provided better reconstruction result compared to the sparse detector placements. Reconstruction error occurred in edges of the congested region with wider detector spacing, lower floating car penetration, and coarser data from AVI. The future extension of this research included better estimation of the model parameters from an automated process or a priori estimate based on Bayesian statistics.

Treiber et al. (2011) expanded the method developed by Treiber & Helbing (2002) to the 'Generalized Adaptive Smoothing Method' to incorporate data from heterogeneous sources (Treiber, Kesting and Wilson, 2011). The motivations were to address the sparseness and noise of a single data source by combining data from multiple sources. The authors used a simulated 12 km highway with 4 loop detectors and 10 floating cars. Using point speed from a few floating cars along with the detector-based speeds, the smoothed velocity reconstruction achieved better accuracy compared to the single source-based reconstructions. Using this method, the authors recommended data reconstruction for places with detectors spacing up to 1.8 mile. The close positioning of detectors was recommended for bottleneck locations to accurately identify the congestion. The generalized smoothing method is used by later studies (Ottaviano, Cui and Chow, 2017).

One extension of the Van Lint and Hoogendoorn (2010) study was performed in Li et al. (2016) where the authors developed the fusion algorithm for urban expressways. The authors modified the weight function to fuse data from multiple sources based on the data captured from the urban expressway. They used real-world data from a 10 km corridor in Beijing. The data collection interval for the loop detectors was 2 min, while for GPS-based vehicles it was 5 min. Using only vehicle data, a minimum 5% penetration rate was identified to provide a reliable travel time estimation. In addition, the fusion of loop detector data and GPS data (at 2% penetration) outperformed the travel time estimation from single data sources.

2.2. DATA FUSION

In one study about data fusion, the authors summarized the issues about input data based on the sensors' setup and operational characteristics, captured data characteristics and influence of the external environment (Khaleghi *et al.*, 2013). They reviewed state-of-the-art data fusion methods based on specific data issues to provide an organized view of data fusion methods, specifying the applicability, advantages, and limitations. For imperfect data issues, the authors found these methods are used: a probabilistic method for data uncertainty, an evidential method for uncertain and ambiguous data, and a rough set-theoretic for data without preliminary information. The stochastic adaptive sensor modeling is used if outliers exist. According to this study, the emerging data fusion methods included opportunistic information fusion model (dynamically discover sensor, computational load, and dynamic fusion rule), and adaptive data fusion (adaptive Kalman filter, reinforcement learning). However, this review paper does not include the discussion on deep learning models for data fusion.

In one study, Liu et al. (2020) discussed the categories of deep learning-based urban multi-source big data fusion, and challenges and methods of dealing with urban big data. They summarized the fusion method (including deep learning) based on spatio-temporal characteristics of the data, and future deep learning-based data fusion research directions. At first, based on the spatio-temporal characteristics of the data, they categorized fusion methods into three groups: feature-level based (using similar features from heterogeneous sources), stage-based (disaggregating the fusion in multiple connected stages), and semantic meaning based (using the similarity and correlation of the multiple data). Later, deep learning-based fusion methods (feature-level fusion) were categorized into three groups, which are: output-based (late fusion), input-based (early fusion), and double stage based (both early and late fusion). The data quality, sparsity, modality, and spatio-temporal characteristics of the data affect big data fusion. The future research directions include the incorporation of deep learning based fusion method to handle missing data and multi-modality in the urban big data, and the inclusion of multi-model fusion (e.g., Convolutional Neural Network or CNN and Recurrent Neural Network or RNN).

In their study, Ambühl & Menendez (2016) used a weight-based model to determine the average flow and density of a network to estimate the macroscopic flow model (MFD). The developed model does not need prior information like Kalman Filter does, only needs the information from probe vehicles and homogeneously situated loop. The motivation of the authors was to develop a simple network-level estimation model. They determined flow and density from detector and probe separately, and estimated probe vehicle penetration based on probe vehicle number and detector captured total vehicles. Using individual flow and density data from loop and probe vehicles, the authors calculated network flow and density given fixed weights and probe penetration. Two simulation networks with a grid system and downtown Zurich. The authors found fusion produces better result compared to loop and probe alone. The measurement errors of both loops and vehicles had no impact of estimation as MFD uses aggregated information, which can nullify the individual errors. However, this method is not applicable for low detector penetration. The authors assumed homogeneous distribution of loop.

Patire et al. (2015) estimated travel time using point-based GPS data from probe and loop detector data. One challenge was to map probe-based GPS points to the freeway. The authors used a Path Inference Filter for the projection. At first, the authors filtered and down selected loop and probe data, then fed these data into the fusion engine. The fusion methods used the Godunov scheme discretized Lighthill and Whitham model to estimate speed. Data were fused using Ensemble Kalman Filtering (EnKF). The authors used an iterative calibration process to fine-tune the model parameters. The EnKF was calibrated with Bluetooth based data. The model was validated with multiple networks. One key result was the relationship between GPS sample-rate and vehicle penetration rate. On freeways, better travel time estimation was achieved using data with a low sample-rate and a high penetration rate, than that achieved using data with a high sample-rate and a low penetration rate. In addition, where no loop detectors existed, travel time could be estimated with reliable accuracy using only probe vehicle data.

In another study, the authors estimated route flow based on loop detector and cellular data fusion using convex optimization (Wu *et al.*, 2015). As urban traffic may not be in any equilibrium state, the authors developed a data driven route flow estimation model without relying on equilibrium-based models. The optimization method fitted the estimated route flow with loop detector data. The inputs to the flow estimation model included road network, origin-destination demands, set of routes, cellular-based flow measurements, and loop detector data from a subset of the corridors from the entire network. Numerical studies were conducted where authors achieved 99% route flow estimation accuracy for the I-210 freeway near Los Angeles with data from loop detector and cellular networks. One limitation of the study is that the authors considered static traffic demands for the network, which is not realistic for real-world implementation.

Wang et al. (2019) reconstructed traffic data from multiple sources. They considered both internal structure of single data and relationships among multi-source data to reconstruct the data. They developed framework separated noise from the real data and measured relationship among multiple data using the fundamental flow diagram relationships. The Alternating Direction of Multipliers optimization method was used to obtain the reconstructed data. The authors used real-world average speed data from 28 links using cell phone and floating car for single parameter reconstruction. For multi-parameter reconstruction, they used real-world speed, occupancy, and volume data collected from a microwave detector. For single-parameter reconstruction, with 80% data loss the reconstruction error was less than 15%. However, this method is only applicable to uninterrupted flow. In addition, the authors' assumed that the single measurements for the same position from multiple sensors are same, which is not always valid in the real-world.

In their study, Wright & Horowitz (2016) estimated freeway density from loop and probe vehicle data using Rao-Blackwellized particle filter to match the solution of traffic partial differential equations with the available sensor data. The assumptions were that the density and velocity of a time are independent of a previous time. The authors used a stochastic cell transmission model (CTM), macroscopic flow model, initial density distribution, per-link predicted velocity distribution, and likelihood functions for density and velocity and performed recursive one-step CTM model update.

3. METHODS

This section discusses the research method used for the performance measure estimation using both the single point detector-based method and the hybrid method. The existing methods of computing VMT, VHT, and VHD are adopted based on the methods used by the Caltrans PeMS. In the end, this section discusses the evaluation setup and scenarios to evaluate the frameworks. Figure 3-4 shows the overall framework of both PeMS and hybrid methods.

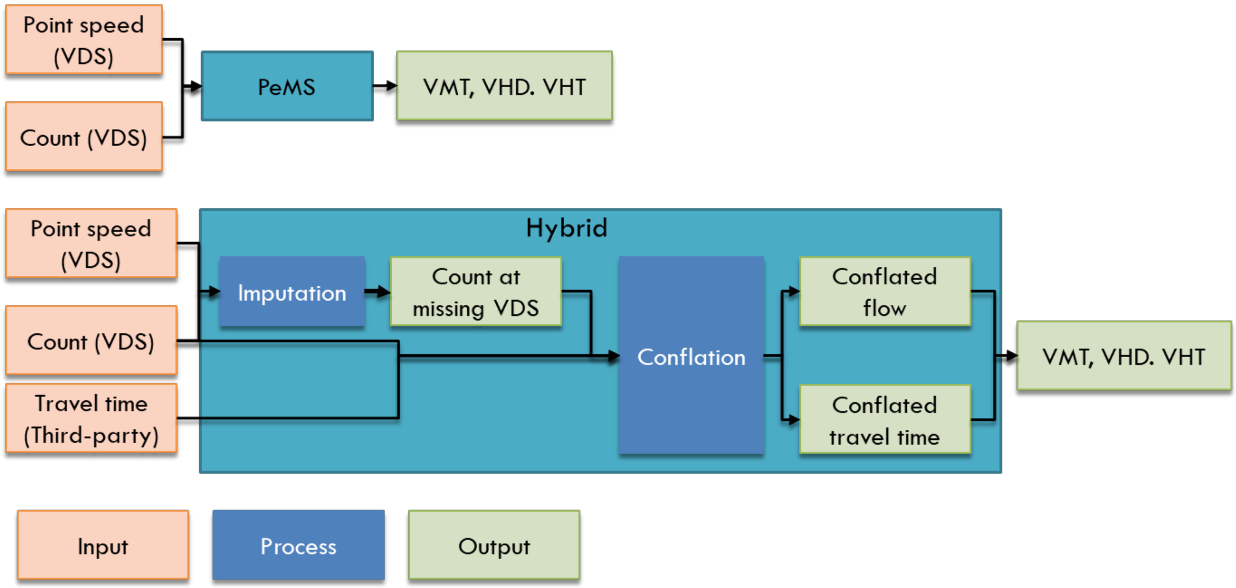


Figure 3-4 Performance measure estimation framework

3.1. PEMS METHOD

This section describes the existing method to calculate performance measures currently employed in PeMS.

3.1.1. PEMS SPEED ESTIMATION

Using single loop detectors, flow (i.e., the number of vehicles passing the detector during a certain time interval) and occupancy (i.e., the percentage of the time during which the detector is occupied) values are reported. To calculate speed from flow and occupancy, another parameter called a g-factor is estimated which is the average length of vehicles passing over the detectors. Assume that for any time interval *i*, o_i, and q_i are the occupancy and flow values, respectively, for the loop detector. Using multiple iterations with the experimental setup (discussed in Section 3.3), the suitable g-factor values for each lane on the freeways are estimated. The preliminary speed, s_i, is:

$$s_i = \frac{g \cdot q_i}{o_i} \qquad \textit{Eq. 3-1}$$

Here *g* is the g-factor value. Using an exponential filter, the final calculated speed estimate v_i is obtained from the estimated speed s_i, as shown in **Eq. 3-2**. The variable, w_i, can be estimated with **Eq. 3-3**. The value of smoothing parameter, *a*, is considered from (Zwet *et al.*, 2003).

$$v_i = w_i \cdot s_i + (1 - w_i) \cdot v_{i-1} \qquad \textit{Eq. 3-2}$$

$$w_i = \frac{q_i}{q_i + a} \qquad \textit{Eq. 3-3}$$

The final calculated speed represents a point-speed for the associated loop detector. At any location, several individual loop detectors can form a VDS as shown in Figure 3-5. To estimate the speed at a VDS for a time interval *i*, estimates from all loop detectors are averaged.

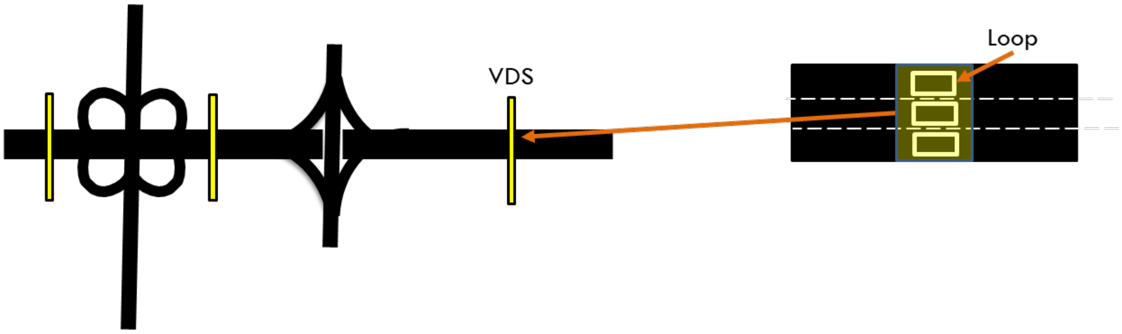


Figure 3-5 Schematic of multiple loops forming single VDS

3.1.2. PEMS PERFORMANCE MEASURES ESTIMATION

PeMS assumes that each VDS is representative of the freeway segment from the upstream midpoint to the downstream midpoint of neighboring VDS. In Figure 3-6, the green double arrow denotes the freeway segment of the VDS that lies within it. The yellow stars mark the midpoint of two successive VDS. In terms of traffic data, the key measurements captured or estimated at the VDS locations are vehicle counts, occupancies, and speeds.

For an interval time interval *i*, VMT is the sum of the total miles driven by all vehicles for a freeway in that time interval. VMT is represented in the units of vehicle-miles and can be calculated over a specific interval by the following equation:

$$\mathrm{VMT} = \sum_i L \cdot q_i \qquad \textit{Eq. 3-4}$$

Where q_i is the number of vehicles that passed over the VDS and L is the length of its associated freeway segment.

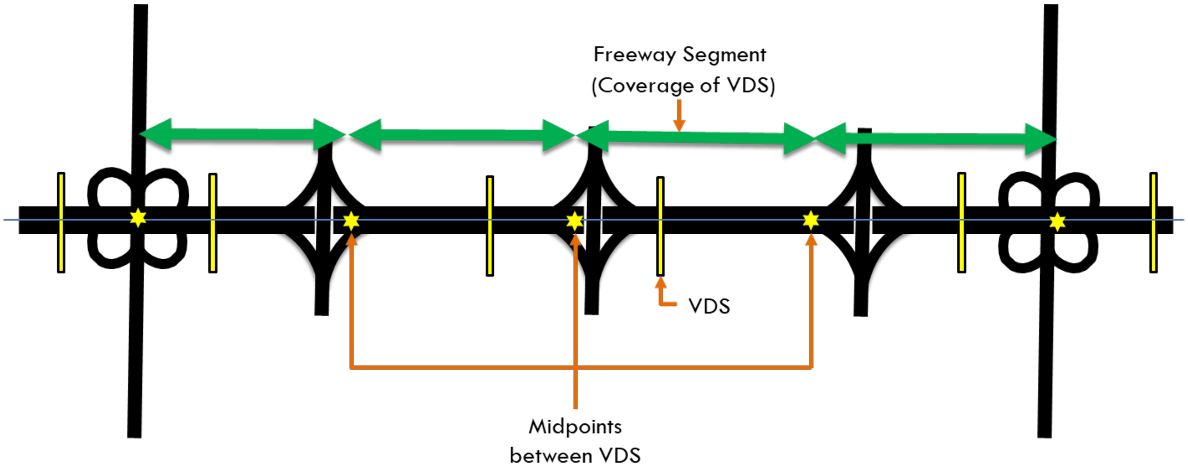


Figure 3-6 Freeway segment representing coverage of VDS

For any interval time *i*, VHT is the sum of the total hours driven by all vehicles for a freeway in that time interval. VHT is represented in the units of vehicle-hours and can be calculated over a specific interval by the following equation.

$$\mathrm{VHT} = \sum_{i} \frac{L \cdot q_i}{v_i} \qquad \textit{Eq. 3-5}$$

Where, v, is the speed from the VDS in question.

Delay is represented in units of vehicle·hours and is calculated against a threshold speed.

$$\mathrm{VHD} = \sum_{i} q_i \left(\frac{L}{v_i} - \frac{L}{b}\right) \qquad \textit{Eq. 3-6}$$

Where, v, is the speed from the VDS in question and, *b*, is the threshold speed. The 65 and 35 mph threshold speeds are considered in this research for freeway mainline and freeway-freeway connectors, respectively.

3.1.3. COMMENTS ON THE MPR METHOD

The two main opportunities for errors to arise in this method are:

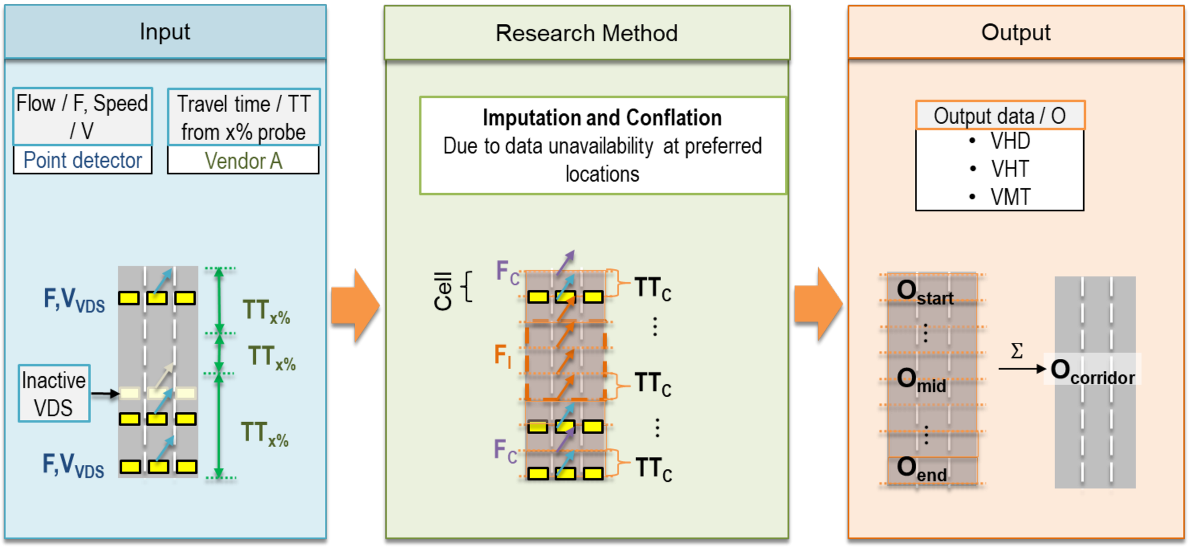
- usage of the g-factor approximation to estimate speeds
- usage of a point measurement to approximate the measurement across an expanse of road

Single loops, as predominantly deployed, do not actually measure speeds. Even if the point speed is measured with other sensors (e.g., dual loops), the point speed does not reflect the overall operational condition for a freeway segment. One may expect that direct measurement of travel-times that are possible with third-party data may be advantageous.

3.2. HYBRID METHOD

The hybrid method is illustrated using Figure 3-7 which shows input data, intermediate analysis steps, and output with the estimated performance measures. The input data include flow and speed from the VDS, and travel time (TT) data provided by third-party vendors. For the analysis of this research, one vendor is considered (Vendor A) who provides data in a separate spatial reference system which does not match that used by Caltrans PeMS. For the freeway mainline, Figure 3-7 shows such a situation where the links with travel time data do not align with the VDS locations. Also, the travel time data is not based on the whole vehicle population, it is from a sample probe vehicle group (x% of the whole population).

Imputation is performed where data is missing in case of any inactive or decommissioned VDS. Once data are available in all VDS, flow is conflated or projected to the desired cells along the freeway. Based on the conflated flow, travel time data from third-party vendors are also conflated. Once both flow and travel time data are available on the desired cells, the final performance measures are calculated by aggregating data from all cells in the freeway. In this technical report, the cell-based calculation is conducted for the freeway-mainlines. However, freeway-freeway connectors are not subdivided into smaller pieces. For the freeway-freeway connectors, the data fusion is not performed using cells, rather it is done for the whole length of the connectors.



VDS = Vehicle Detector Station
C = Conflated
I = Imputed

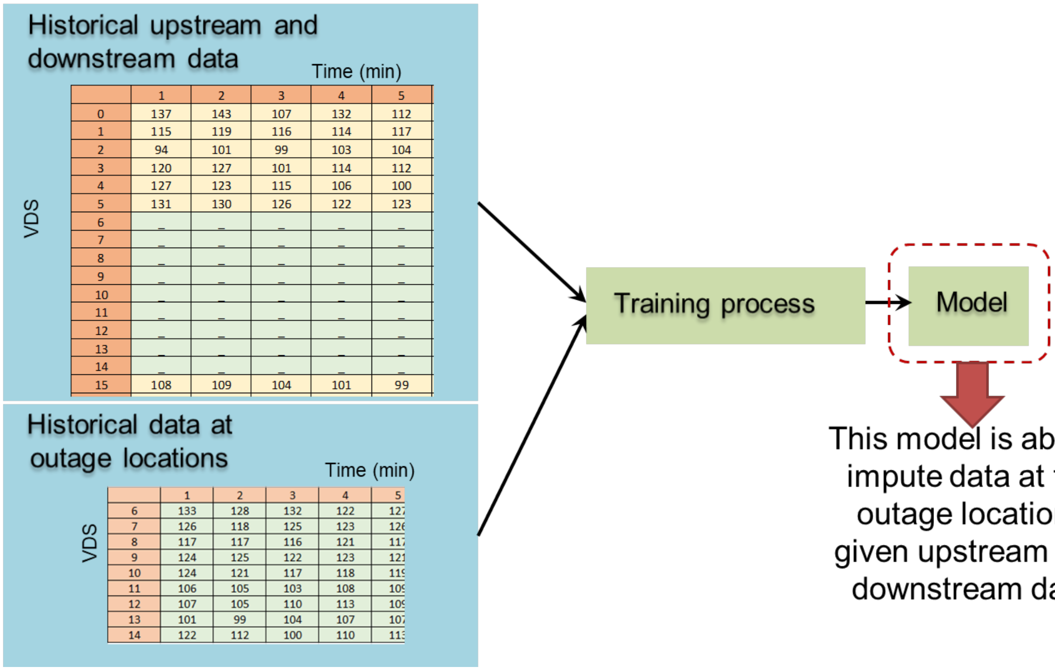
Figure 3-7 Hybrid method to estimate performance measures

3.2.1. INCOMPLETE VDS DATA IMPUTATION

In this analysis, a machine learning (ML)-based imputation method is studied to impute flow in the missing VDS locations. At first, two images are generated to train the ML model as shown in Figure 3-8. The first image is generated with the flow data from the VDS locations that includes both VDS with and without missing data. Here VDS with missing data does not have any data at the missing locations. The second image has data of the missing locations. The x-axis of the image represents the per minute time interval, and the y-axis represents the VDS. For a total of *B* minutes, the following matrix *C* represents the input data for freeway with *A* VDS, or equivalently, *A* freeway segments.

$$\mathrm{C} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,B} \\ \vdots & \ddots & \vdots \\ c_{A,1} & \cdots & c_{A,B} \end{bmatrix} \qquad \textit{Eq. 3-7}$$

Where *c* represents the flow at VDS, *a,* for the *b*-th time interval (in minutes). The motivation of having an image-based analysis is to use both spatio-temporal information to impute missing data (Ma *et al.*, 2017). A regular image can have pixel values ranging from 0 to 255, whereas in this study the pixel values are flow values at a certain location and time. Also, regular images can have 1 channel (grayscale image) or 3 channels (Red Green Blue image). In this analysis, only one channel is used for the flow data. The captured feature maps by different layers of the ML models are the relationship between traffic flow and VDS locations.

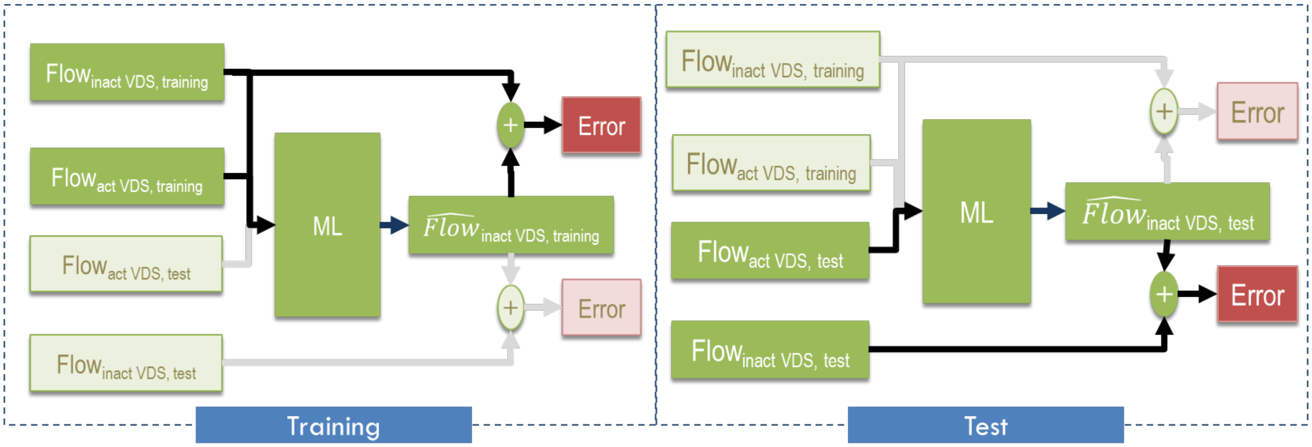

Historical upstream and downstream data
Time (min)
VDS
Historical data at outage locations
Time (min)
VDS
Training process
Model
This model is able to impute data at the outage locations given upstream and downstream data

	1	2	3	4	5
0	137	143	107	132	112
1	115	119	116	114	117
2	94	101	99	103	104
3	120	127	101	114	112
4	127	123	115	106	100
5	131	130	126	122	123
6	_	_	_	_	_
7	_	_	_	_	_
8	_	_	_	_	_
9	_	_	_	_	_
10	_	_	_	_	_
11	_	_	_	_	_
12	_	_	_	_	_
13	_	_	_	_	_
14	_	_	_	_	_
15	108	109	104	101	99

	1	2	3	4	5
6	133	128	132	122	127
7	126	118	125	123	126
8	117	117	116	121	117
9	124	125	122	123	121
10	124	121	117	118	119
11	106	105	103	108	109
12	107	105	110	113	109
13	101	99	104	107	107
14	122	112	100	110	113

Figure 3-8 schematic of imputation model development

Figure 3-9 shows the two phases related to the ML models, which are training and test. Different datasets are used for training and test. At the training phase, the model hyperparameters are selected based on the cross-validation error using the training dataset. Hyperparameters are those variables which affect the performance of the model, and that cannot be estimated from the data by the model itself. For CNN, these hyperparameters are the number of hidden layers, epoch (number of times the data is passed forward and backward through the network), batch size (number of data present in a group that is passed through the model), learning rate (the rate at which the weights are updated), and decay rate (the rate at which learning rate changes). Later the final trained model's performance is measured based on the estimated flow and actual flow on the test dataset. The following subsections further discuss the imputation methods.



$Flow_{inact\ VDS,\ training}$ = Flow values (veh-min) of the inactive VDSs in the training scenario
$Flow_{inact\ VDS,\ test}$ = Flow values (veh-min) of the inactive VDSs in the test scenario
$Flow_{act\ VDS,\ training}$ = Flow values (veh-min) of the active VDSs in the training scenario
$Flow_{act\ VDS,\ test}$ = Flow values (veh-min) of the active VDSs in the test scenario
$\widehat{Flow}_{inact\ VDS}$ = Estimated flow values of the inactive VDSs

Figure 3-9 ML-based training and test

CONVOLUTIONAL NEURAL NETWORK

CNN is a widely accepted ML technique for image-based analysis. Here, a brief discussion is included to explain the underlying concept of the CNN model, which is implemented based on (Ma *et al.*, 2017). In this model, convolution and pooling layers are the core parts as shown in Figure 3-10. In the convolution layer, features from the input image are extracted by sliding a filter over the image. A convolutional filter's dimension is height X width X channel, where channel number depends on the channel of the input image. In the pooling layer, the dimensions of the incoming data are reduced to minimize the number of parameters. For a certain layer *l* (with a total number of convolution filters, *t*), assume that the input, output, weight, bias, and channel index of convolutional filters are *c*, *y*, *p*, *j*, and *m*, respectively. The following shows the output from the first convolutional and pooling layer.

$$y_1^m = \mathrm{pool}\left(\beta \cdot (p_1^m \cdot c_1^m + j_1^m)\right), m \in [1, t_l] \qquad \textit{Eq. 3-8}$$

Here β is the activation function. The following equation shows the output from the additional convolutional and pooling layers.

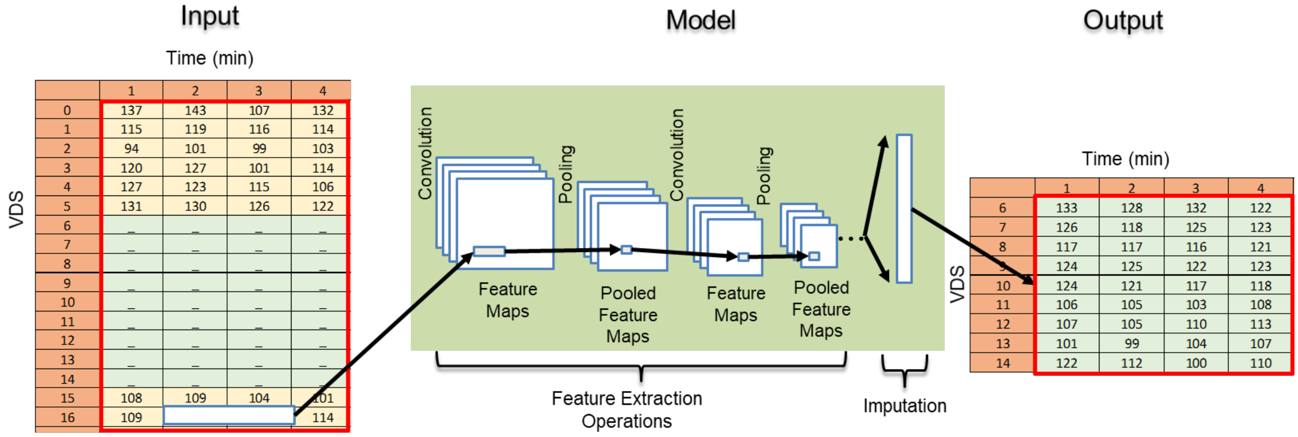
$$y_1^m = \mathrm{pool}\left(\beta \cdot \left(\sum_{n=1}^{t_{l-1}} (p_l^m \cdot c_l^n + j_1^m)\right)\right), m \in [1, t_l] \qquad \textit{Eq. 3-9}$$

A dense layer concentrates (in other word, flattens) the final features learned by the intermediate layers, which can be written as the following equation. Here the depth of CNN is denoted by L.

$$y_L^{\mathrm{flatten}} = \mathrm{flatten}\left([y_L^1, y_L^2, \ldots, y_L^m]\right), m = t_L \qquad \textit{Eq. 3-10}$$

In the end, the model output ($\hat{y}$) (as shown in **Eq. 3-11**) is generated with a fully connected layer, *f*.

$$\hat{y} = p_f \cdot y_L^{flatten} + j_f \qquad \textit{Eq. 3-11}$$



VDS	1	2	3	4
0	137	143	107	132
1	115	119	116	114
2	94	101	99	103
3	120	127	101	114
4	127	123	115	106
5	131	130	126	122
6	_	_	_	_
7	_	_	_	_
8	_	_	_	_
9	_	_	_	_
10	_	_	_	_
11	_	_	_	_
12	_	_	_	_
13	_	_	_	_
14	_	_	_	_
15	108	109	104	101
16	109			114

VDS	1	2	3	4
6	133	128	132	122
7	126	118	125	123
8	117	117	116	121
9	124	125	122	123
10	124	121	117	118
11	106	105	103	108
12	107	105	110	113
13	101	99	104	107
14	122	112	100	110

Figure 3-10 CNN-based imputation model

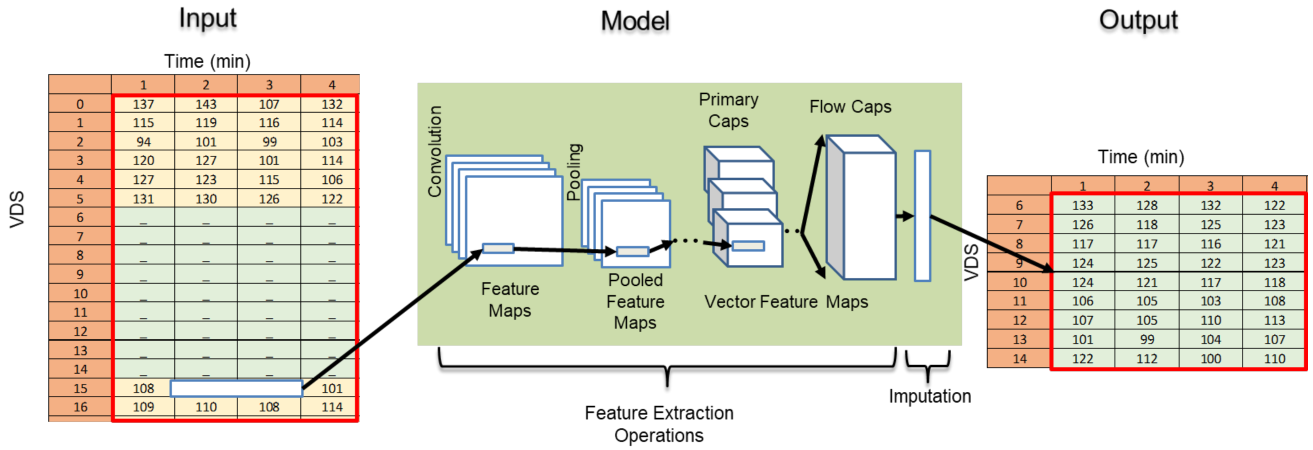
The model output, $\hat{y}$, is the imputed data at the VDS locations with the missing data. In the training phase, the model hyperparameters are optimized, and later the trained model is used with the test dataset.

CAPSULE NETWORK

There is a chance of losing information in the pooling function operated by CNN, which mainly down samples the incoming data from the input image or previous layer. To address the issue, capsule network (CapsNet) is developed. CapsNet has layers with capsules (a group of neurons) with additional information handling capabilities. Layers of CapsNet store information about different properties of the same object. There are two main capsule layers, the first one is the primary capsule layer and the second one is the flow capsule layer, as shown in Figure 3-11. The input and output of capsule layers have vector forms, while regular neurons in CNN have scalar forms. Assume *i* and *j* are two capsules at the upper and lower levels. The variables ∝, *c, j*, and *y* are the non-linear activation function, coupling coefficient, weight matrix, and output vectors, respectively. **Eq. 3-12** shows the output vector from a capsule

$$y_j = \propto (\sum_i c_{ij} \cdot j_{ij} \cdot y_i) \qquad \textit{Eq. 3-12}$$

The coupling coefficient, *c* is estimated using a dynamic routing method (Sabour, Frosst and Hinton, 2017) which basically specifies if the weights estimated from the low-level capsules are in agreement with the weights estimated by the high-level capsules.



Time (min)

VDS	1	2	3	4
0	137	143	107	132
1	115	119	116	114
2	94	101	99	103
3	120	127	101	114
4	127	123	115	106
5	131	130	126	122
6	_	_	_	_
7	_	_	_	_
8	_	_	_	_
9	_	_	_	_
10	_	_	_	_
11	_	_	_	_
12	_	_	_	_
13	_	_	_	_
14	_	_	_	_
15	108			101
16	109	110	108	114

Time (min)

VDS	1	2	3	4
6	133	128	132	122
7	126	118	125	123
8	117	117	116	121
9	124	125	122	123
10	124	121	117	118
11	106	105	103	108
12	107	105	110	113
13	101	99	104	107
14	122	112	100	110

Figure 3-11 CapsNet-based imputation model

3.2.2. HYBRID DATA CONFLATION

In this subsection, the data conflation method is discussed. The main purpose of this step is to make both flow and travel time data from multiple data sources available on a single spatial reference scheme, in this case, cells along the freeway.

DESIRED CELLS OF ANALYSIS

Figure 3-12 shows a schematic of evenly sized cells along the freeway mainline. In general, travel time data on links do not line up with VDS data on segments, which do not line up with the cells along the desired domain of analysis. The cells are the blue bounding boxes, and they cover freeway mainlines. The motivation of cell-based analysis is to narrow down the locations of the bottlenecks and compute delay properly using conflated flow in each cell. With large cells, variations in flow cannot be properly captured and it can lead to erroneous delay calculation. In this analysis, the cell length is 0.25 miles. Therefore, flow data from the VDS are projected every 0.25 miles. If any VDS is within 200 ft from a cell boundary, the cell location is not considered and the raw VDS data is used as is. At the end of the conflation process, flow data is available at each VDS (measured flow) and cell boundary (conflated flow).

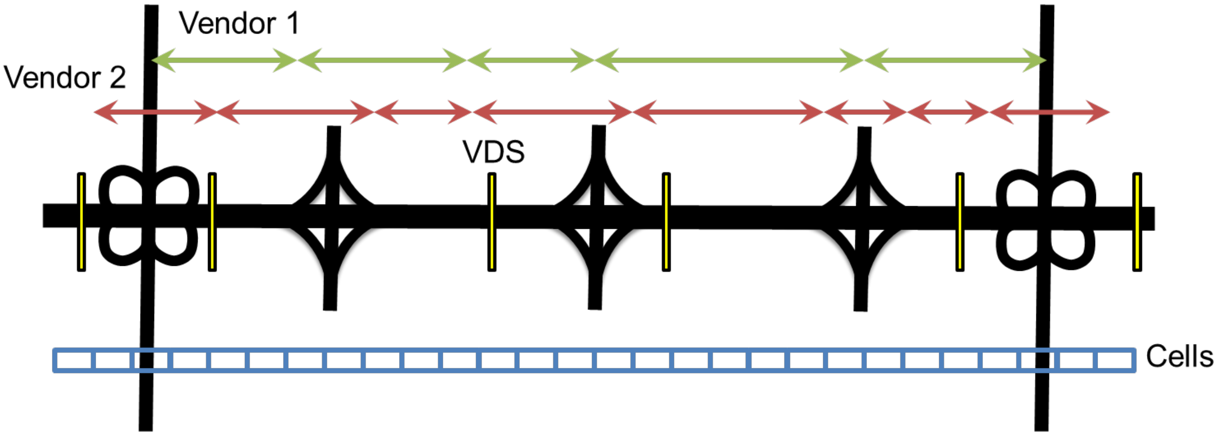


Figure 3-12 Freeway with cells

FLOW CONFLATION

The flow conflation method is developed based on the Generalized Adaptive Smoothing Method (GASM) (Treiber and Helbing, 2002; Treiber, Kesting and Wilson, 2011). The purpose of the method is to spatio-temporarily reconstruct traffic data at specific locations using data captured by point detectors or VDS. At certain cell points (where *x* and *t* are the position and time) on a spatio-temporal domain, the flow data *q* can be calculated using GASM. In this method, $f_{(x,t)}$ is a normalization factor, and *k* is the kernel value. According to GASM, the conflated flow at cell point *(x, t)* is obtained from all VDS captured flow values in the upstream and downstream regions. In GASM, localized smoothing is performed, meaning flow at a certain cell point *(x, t)* is affected strongly by the closer VDS, and weakly by the distant VDS. The widths of spatial and temporal smoothing are δ and μ, respectively.

GASM is developed to overcome the challenge of isotropic smoothing of traffic data (i.e., non-skewing smoothing). The equations of the GASM method are provided here with the non-skewing smoothing. Here *i* and *j* refer to the time interval and VDS number, respectively, for the study period and analysis area. At any cell point located at *x* position, $q_{c(x,t)}$ is the flow at time interval *t*. The variable $\mathrm{q}_{\mathrm{vds(i,j)}}$ is the flow of the *j*-th VDS at the *i*-th time interval.

$$\mathrm{q}_{\mathrm{c(x,t)}} = \frac{1}{\mathrm{f}_{\mathrm{(x,t)}}}\sum_{i=1}^{T}\sum_{j=1}^{N}\mathrm{k}_{(\mathrm{x}-\mathrm{x}_\mathrm{j},\ \mathrm{t}-\mathrm{t}_\mathrm{i})}\cdot \mathrm{q}_{\mathrm{vds(i,j)}} \qquad Eq.\ 3\text{-}13$$

$$\mathrm{f}_{\mathrm{(x,t)}} = \sum_{i=1}^{T}\sum_{j=1}^{N}\mathrm{k}_{(\mathrm{x}-\mathrm{x}_\mathrm{j},\ \mathrm{t}-\mathrm{t}_\mathrm{i})} \qquad Eq.\ 3\text{-}14$$

$$\mathrm{k}_{(\mathrm{x}-\mathrm{x}_\mathrm{j},\ \mathrm{t}-\mathrm{t}_\mathrm{i})} = \exp\left[-\left(\frac{|\mathrm{x}-\mathrm{x}_\mathrm{j}|}{\delta}+\frac{|\mathrm{t}-\mathrm{t}_\mathrm{i}|}{\mu}\right)\right] \qquad Eq.\ 3\text{-}15$$

Here *N* is the total number of VDS, and *T* is the last time interval. GASM includes the idea of skewed smoothing of traffic data. In the free-flow *(ff)* direction, the smoothing is performed with the free-flow propagation speed *(v_{ff})*. In the congested *(cong)* direction, the smoothing is performed with the backward propagation speed *(v_{cong})*. In this analysis, the available VDS are confined to the immediate upstream and downstream VDS, and thus the smoothing method is named as 'Confined Generalized Adaptive Smoothing Method (C-GASM)'. Figure 3-13 shows the situation where intermediate cells between two VDS are conflated with the data from the immediate upstream and downstream VDS.

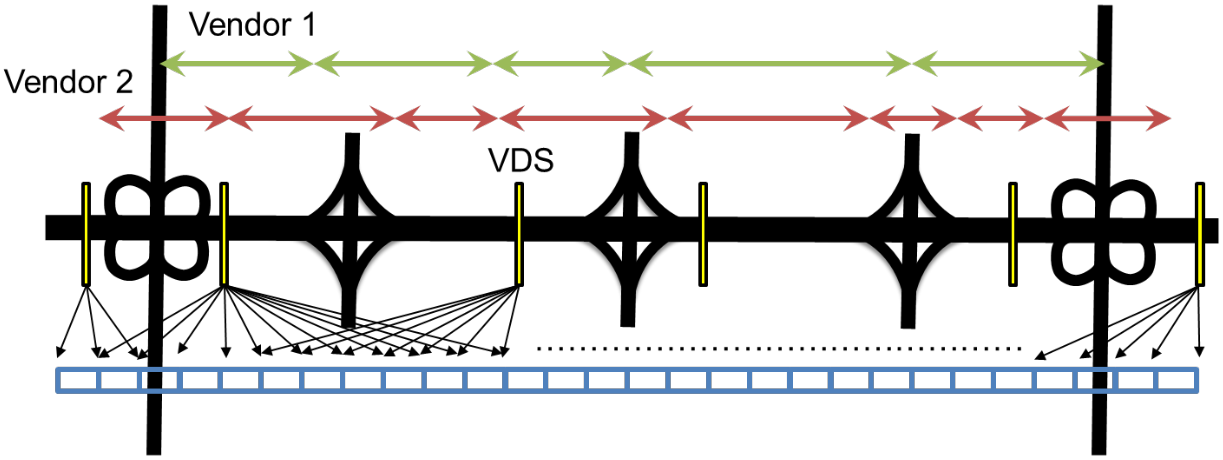


Figure 3-13 C-GASM for conflating with surrounding VDS

Based on the confinement rule, the values for conflated flow in congested and free-flow conditions at cell point *(x, t)* are estimated with **Eq. 3-16** and **Eq. 3-18**, *r*espectively. Here *u* is the upstream VDS while *d* is the downstream one.

$$q_{c,ff(x,t)} = \frac{1}{f_{ff(x,t)}} \sum_{i=1}^{T} \left(k_{\left(x-x_u,t-t_i-\frac{x-x_u}{v_{ff}}\right)} \cdot q_{vds(i,u)} + k_{\left(x-x_d,t-t_i-\frac{x-x_d}{v_{ff}}\right)} \cdot q_{vds(i,d)}\right) \qquad Eq.\ 3\text{-}16$$

$$f_{ff(x,t)} = \sum_{i=1}^{T} \left(k_{\left(x-x_u,t-t_i-\frac{x-x_u}{v_{ff}}\right)} + k_{\left(x-x_d,t-t_i-\frac{x-x_d}{v_{ff}}\right)}\right) \qquad Eq.\ 3\text{-}17$$

$$q_{c,cong(x,t)} = \frac{1}{f_{cong(x,t)}} \sum_{i=1}^{T} \left(k_{\left(x-x_u,t-t_i-\frac{x-x_u}{v_{cong}}\right)} \cdot q_{vds(i,u)} + k_{\left(x-x_d,t-t_i-\frac{x-x_d}{v_{cong}}\right)} \cdot q_{vds(i,d)}\right) \qquad Eq.\ 3\text{-}18$$

$$f_{cong(x,t)} = \sum_{i=1}^{T} \left(k_{\left(x-x_u,t-t_i-\frac{x-x_u}{v_{cong}}\right)} + k_{\left(x-x_d,t-t_i-\frac{x-x_d}{v_{cong}}\right)}\right) \qquad Eq.\ 3\text{-}19$$

To calculate a single smoothed flow value for *(x, t)*, a weighted filter is used. With **Eq. 3-20**, the final flow value at the cell point *(x,t)*, which is $q_{f(x,t)}$, is estimated.

$$q_{f(x,t)} = z_{(x,t)} \cdot q_{c,cong(x,t)} + (1 - z_{(x,t)}) \cdot q_{c,ff(x,t)} \qquad Eq.\ 3\text{-}20$$

The weight $z_{(x,t)}$ is calculated with an s-shape function, which depends on crossover speed (v_{cr}) and transition width (Δv) from congestion to free flow.

$$z_{(x,t)} = \frac{1}{2} \cdot [1 + \tanh(\frac{v_{cr} - \min(v_{c,ff(x,t)}, v_{c,cong(x,t)})}{\Delta v})] \qquad Eq.\ 3\text{-}21$$

At the cell point *(x,t)*, values of $v_{c,ff(x,t)}$ and $v_{c,cong(x,t)}$ are calculated using equations similar to Eq. 3-16 and Eq. 3-18 with the VDS-captured speeds. C-GASM based flow conflation method depends on parameters such as $\delta, \mu, v_{ff}, v_{cong}, v_{cr}$, and Δv. The typical values of these parameters are discussed in(Treiber and Helbing, 2002; Treiber, Kesting and Wilson, 2011). In this analysis, these parameters are selected based on multiple trials where the chosen set of acceptable values gives the highest accuracy. Figure 3-14 shows an overview of the flow conflation method using C-GASM.

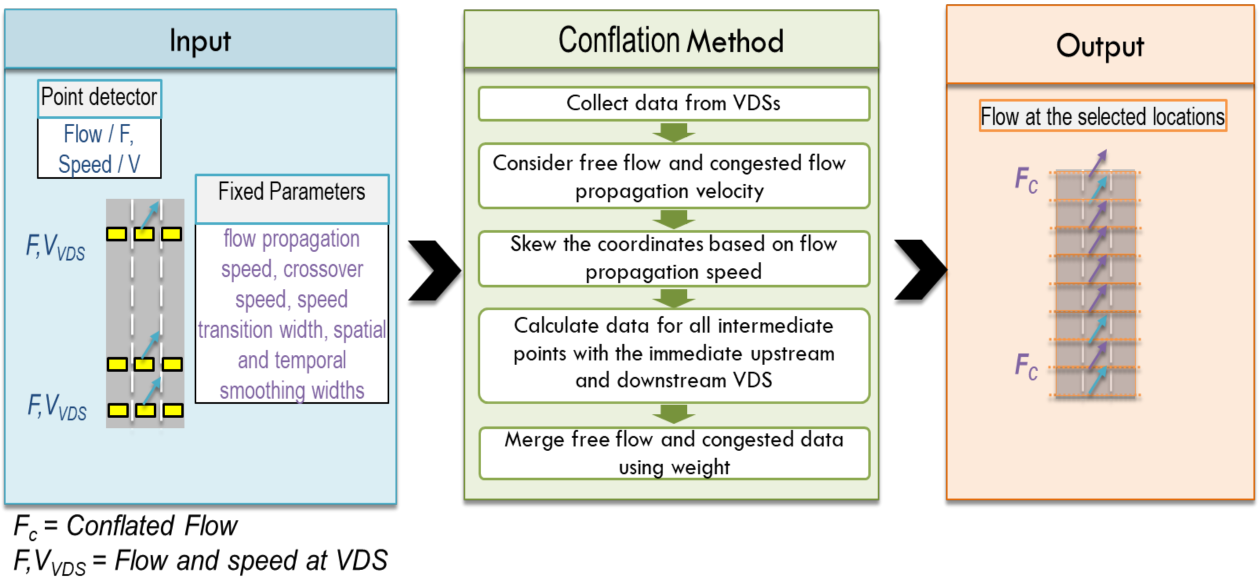

Input
Point detector
Flow / F,
Speed / V
Fixed Parameters
flow propagation speed, crossover speed, speed transition width, spatial and temporal smoothing widths
F,V_VDS
F,V_VDS
Conflation Method
Collect data from VDSs
Consider free flow and congested flow propagation velocity
Skew the coordinates based on flow propagation speed
Calculate data for all intermediate points with the immediate upstream and downstream VDS
Merge free flow and congested data using weight
Output
Flow at the selected locations
F_C
F_C
F_c = Conflated Flow
F,V_VDS = Flow and speed at VDS

Figure 3-14 C-GASM method

THIRD-PARTY DATA CONFLATION

Once the flow values from the point sensors are conflated to the desired cell locations, the next step is to conflate the third-party data on the same cells. Travel time data provided by a third party is conflated to the desired cells, as shown in Figure 3-15, where the link is divided by the overlapping cells. The assumption of travel time conflation is that travel time data can be distributed along the links to the cells based on the vehicle number distribution in the cells. The higher number of vehicles in a cell will result in higher travel times, and vice versa. Assume that, vendor A provided travel time data for a link is TT_i at time interval *i*. In that link, the total number of cells is *G* which divide the link into *(G+1)* parts. For a certain cell located at *x* on that link, the associated travel time data *($tt_{x,i}$)* at time interval *i* from vendor A, is estimated with this equation.

$$tt_{x,i} = TT_i \cdot \frac{c_{x,i}}{\sum_{n=1}^{G+1} c_{n,i}} \qquad Eq.\ 3\text{-}22$$

The number of vehicles in a cell ($c_{x,i}$) located at *x* for the time interval *i* can be estimated by the conflated density on the cell, and associated length of the link from that cell to the next cell. The conflated density is estimated with the C-GASM method.

Travel time for the cells, which cover edges of multiple links, are calculated by aggregating travel time for those link edges at time interval *i*. Travel time data from multiple vendors can be estimated with a weighted sum approach. The weight (∅) can be assigned based on the confidence of the third-party vendor provided data. The confidence can be related to the travel time data characteristics (penetration level of probe vehicles, real-time data availability) of the vendor provide data.

$$tt_{x,i} = \emptyset_A \cdot tt_{A,x,i} + \emptyset_B \cdot tt_{B,x,i}, \emptyset \in [0,1], \emptyset_A + \emptyset_B = 1 \qquad Eq.\ \ 3\text{-}23$$

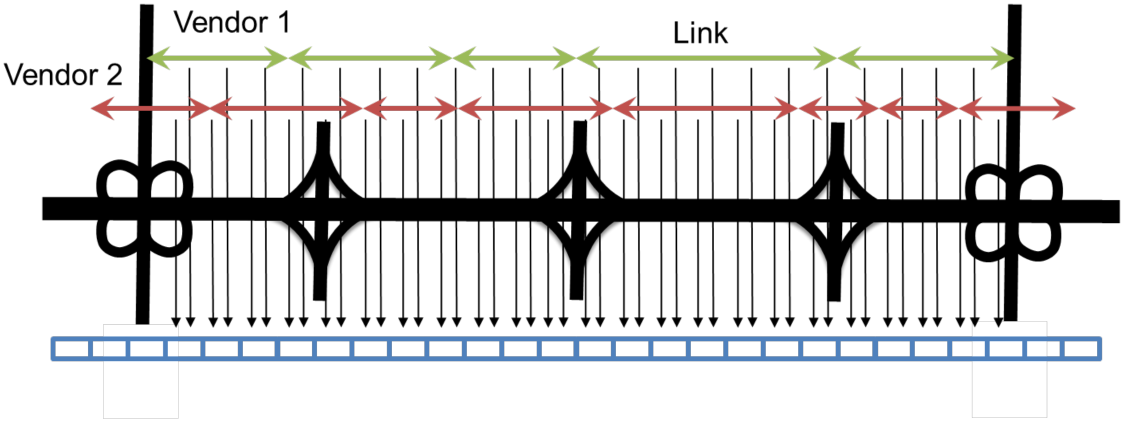


Figure 3-15 Travel time conflation

If multiple vendors have same weights, ∅ will have equal values, and the sum of all weights will be equal to 1. In this research evaluation, only one vendor is considered.

3.2.3. HYBRID PERFORMANCE MEASURES ESTIMATION

Once multiple data from different sources are conflated on the same network along the desired cells, VMT, VHT, and VHD are calculated for each cell. Final values for the freeway are calculated by summing up the values for the individual cell, as shown in Figure 3-7.

3.3. EXPERIMENTAL SETUP

In order to evaluate the performance of the PeMS and hybrid methods, an experimental setup is developed and used with a simulated model of the I-210 corridor.

3.3.1. SIMULATION MODEL

Figure 3-16 shows the calibrated model of the I-210 corridor that is used in this research. The simulation model is developed for the Connected Corridors program, which has different roadways (freeways, ramps, and arterials) calibrated for both weekends and weekdays (Connected Corridors, 2020).

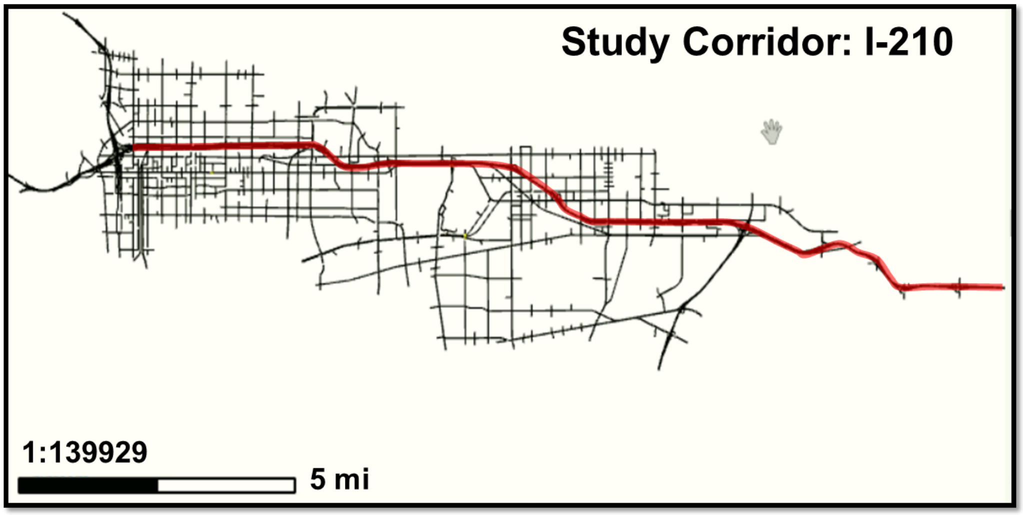


Figure 3-16 I-210 simulation model

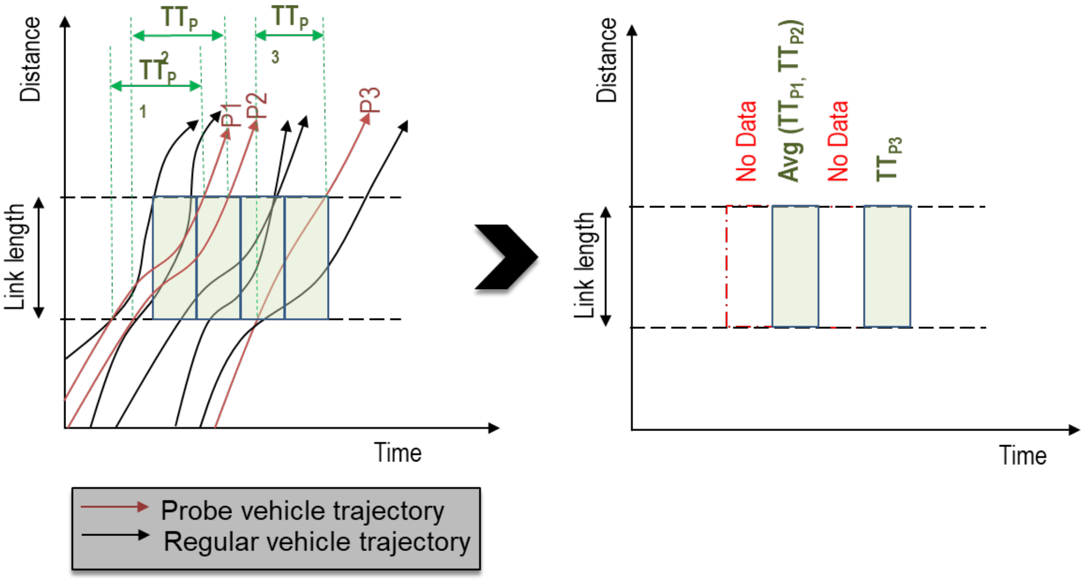
The red highlighted freeway used in this analysis is the westbound portion of the I-210 freeway. The VDS in the simulated model are laid out following the VDS placement on the real-world freeway. To synthesize the third-party vendor travel time data, raw location data is collected from a sample of the total vehicles (e.g., 5% of the simulated vehicles). The probe vehicle data is only considered when the vehicle data is available for the initial and last 10% part of the link. This is done to ensure that the vehicle has crossed the link. From the initial and final location and timestamp data of the associated probe vehicle, travel time is calculated for that vehicle. Later, the data is aggregated for every minute time interval, and the final dataset has the travel time data aggregated for every minute.

For this simulation, only the westbound lanes along a 16-mile portion (highlighted in red) of I-210 are used to calculate VHD for the following scenarios:

1) Before morning peak (6 am - 7 am)
2) Morning peak (7 am - 8 am)
3) Noon time (1 pm - 2 pm)
4) Afternoon peak (5 pm - 6 pm)

There are 33 VDS locations along the 16-mile I-210 corridor (highlighted in red). For conflation and hybrid data fusion, the corridor was divided into 55 cells. The simulated point-sensor data was extracted using the VDS locations, and the simulated third-party vendor data was extracted using an assumed 5% probe penetration of the vehicle population. The findings presented are based on the average of two replications (i.e., different simulation runs using a different random seed) for each of the scenarios.

Figure 3-17 shows the representation of how trajectories are used to calculate travel times. The probe vehicle data from P1, P2, and P3 are captured only when these vehicles crossed the section.



$TT_{x\%}$ = Travel time of x% vehicle (e.g., 5%)
TT_{P1} = Travel time for probe vehicle 1

Figure 3-17 Probe vehicle travel time data generation

3.3.2. SIMULATION SCENARIOS

The following subsections discuss the considerations of different simulation scenarios.

MISSING VDS DATA IMPUTATION

To test the missing data imputation framework using ML-based methods, a real-world PeMS dataset is used. The captured data is obtained from the historic PeMS data warehouse, and only weekday data are used. Figure 3-18 shows the area where data from 9 consecutive VDS are missing on a day. In order to impute the data in the missing region, 7 upstream and downstream VDS for that time interval are used. Both CNN and CapsNet are trained with the same training dataset and evaluated against the same test dataset. The total number of days for this experiment is 55, and data is used for the afternoon peak period (5 pm-6 pm) during weekdays. Among these days, 32 days are used to train the imputation model, and 23 days to evaluate.

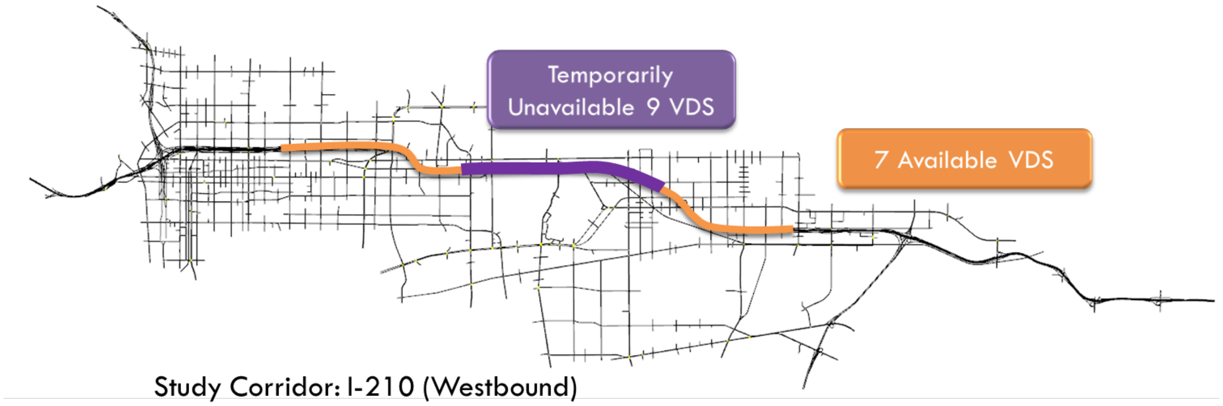

Temporarily Unavailable 9 VDS
7 Available VDS
Study Corridor: I-210 (Westbound)

Figure 3-18 Imputation study area

The details of the CNN-based imputation model are provided in Table 3-1. In the imputation model, there are three convolution + pooling layers. The initial learning rate and decay rate for the model are found to be 0.1, and 0.99, respectively. The batch size is equal to one individual day of data. The model is implemented with the Tensorflow library. The final output of the fully connected layer provides one-minute data for the total 60-minute interval, and for the 9 missing VDS.

Table 3-1 CNN-based imputation model

Layer name	Layer detail	Layer output shape
Convolution Layer #1	Filter =256, Kernel size= 2x2, Activation = 'selu', Padding= 'Same'	22 x 60 x 256 (Total active and inactive VDS x Total time interval in min x Filter size)
Pooling Layer #1	Kernel size = 2x2, Stride=2, Padding='Valid'	11 x 30 x 256
Convolution Layer #2	Filter =128, Kernel size= 3x3, Activation = 'relu', Padding= 'Same'	11 x 30 x 128
Pooling Layer #2	Kernel_size = 2x2, stride=2, Padding='VALID'	5 x 15 x 128
Convolution Layer #3	Filter =32, Kernel size= 3x3, Activation = 'relu', Padding= 'Same'	5 x 15 x 32
Pooling Layer #3	Kernel_size = 2x2, stride=2, Padding='VALID'	2 x 7 x 32
Fully Connected Layer		540 (Total inactive VDS x Total time interval in min)

In the CapsNet-based imputation model, the number of iterations for the dynamic routing process (between Primary Caps and Flow Caps) is limited to 3 iterations. The capsule in the Primary Caps layer represents the VDS, and the dynamic routing algorithm extracts the relationship between all the VDS in the Primary Caps layer. The square-root of the sum of each 16-dimensional capsule vector (squared element) in the Flow Caps layer represents the flow in the VDS. The following Table 3-2 shows the details of the CapsNet-based imputation model. The learning rate and decay rate are 0.05, and 0.9, respectively.

Table 3-2 CapsNet-based imputation model

Layer name	Layer detail	Layer output shape
Convolution Layer #1	Filter =32, Kernel size= 3x3, Activation = 'relu'	22 x 60 x 32 (Total active and inactive VDS x Total time interval in min x Filter size)
Primary Capsule	Capsule output vector size=8D, Filter = 64, Kernel size= 3x3, Activation = 'relu'	10560 x 8 (((Total active and inactive VDS x Total time interval in min x Filter size)/Output capsule dimension) x Output capsule dimension)
Flow Capsule	Capsule output vector size=16D	540 x 16 ((Total inactive VDS x Total time interval in min) x Output capsule dimension)

FREEWAY MAINLINE PERFORMANCE MEASURES

The 16-mile length of the westbound I-210 mainline lanes is used for the evaluation of the PeMS and hybrid methods. Here, no imputation testing is conducted. Along the selected freeway, 33 VDS are available. The final number of cells along the freeway is 55.

VHT and VHD are calculated with respect to the 65-mph speed threshold. VMT, VHT, and VHD are calculated for the following scenarios:

5) Before morning peak (6 am - 7 am)
6) Morning peak (7 am - 8 am)
7) Noon time (1 pm - 2 pm)
8) Afternoon peak (5 pm - 6 pm)
9) Night off-peak (8 pm - 9 pm)

The findings presented in this analysis are based on the average of two replications (i.e., the different simulation runs using different random seeds) for each of the scenarios. For mainline analysis, probe data is provided by 5% of the vehicles.

FREEWAY-FREEWAY CONNECTOR PERFORMANCE MEASURES

In this evaluation, the hybrid calculation is performed for the freeway-freeway connectors using data from both data sources. However, data is not conflated here. For the specific connector, the VDS flow data is used along with the travel time data for that connector. Figure 3-19 shows the connectors and VDS along the study area. VHD and VHT are calculated with respect to the 35-mph speed threshold. Like the mainline analysis, results are presented based on the average of two replications and for five scenarios. For connector analysis, probe data is provided by 100% of the vehicles.

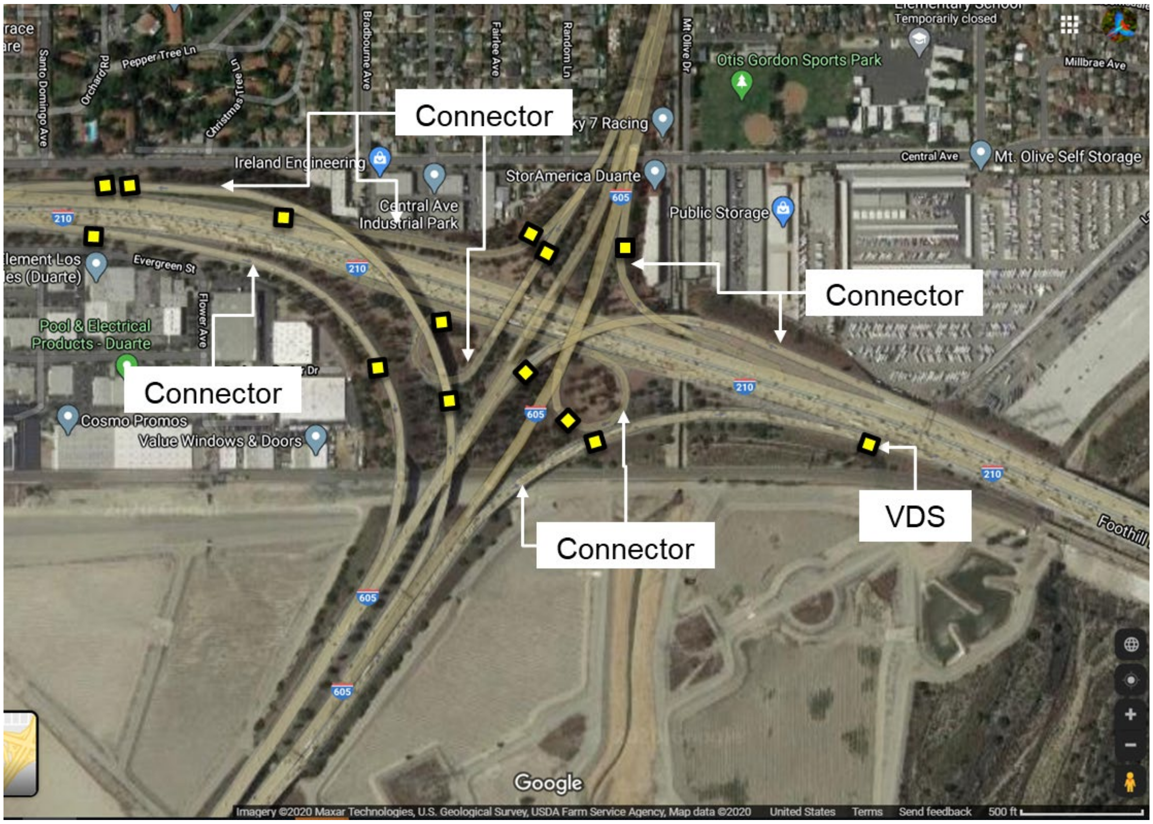


Figure 3-19 I 210 - I 605 connectors

4. RESULTS

In this section, findings from the imputation, conflation, and hybrid data fusion methods are discussed. Although imputation is a part of the hybrid method, in this analysis imputation is not included in the hybrid performance measure estimation. Rather, the findings from the ML-based imputation are discussed separately. The hybrid method is conducted with conflation and data fusion only.

4.1. RESEARCH CONSIDERATION

There are several assumptions considered in the research. For the input data, a one-minute time window is considered. This means data from VDS and a third party are sampled at intervals of one minute. This time interval can be re-sampled to any other preferred time interval. In the analysis, travel time from a third party is used for freeway mainlines; as in reality this data can include data from vehicles on HOV lanes too. The cell-based analysis is conducted for the freeway mainline lanes, whereas for the freeway-freeway connectors, the analysis is conducted for the entire length of the connector.

4.2. ML-BASED MISSING VDS DATA IMPUTATION

ML-based models are data driven models, which means the overall model architecture and model hyperparameters depend on the underlying data. As discussed in Section 3.2.1, the hyperparameters required by ML-based models are estimated by the trial and error method using the training dataset. The first decision to make is to define the required size of the training dataset for each of the ML-based models. Figure 3-20 shows the reduction in error with respect to the increase in the number of training datasets. Here real-world data from the PeMS website was used in the model. As shown in the figure, the performance of both CNN and CapsNet becomes almost steady after 32 datasets. For this reason, data from 32 days are used as the training dataset for both models.

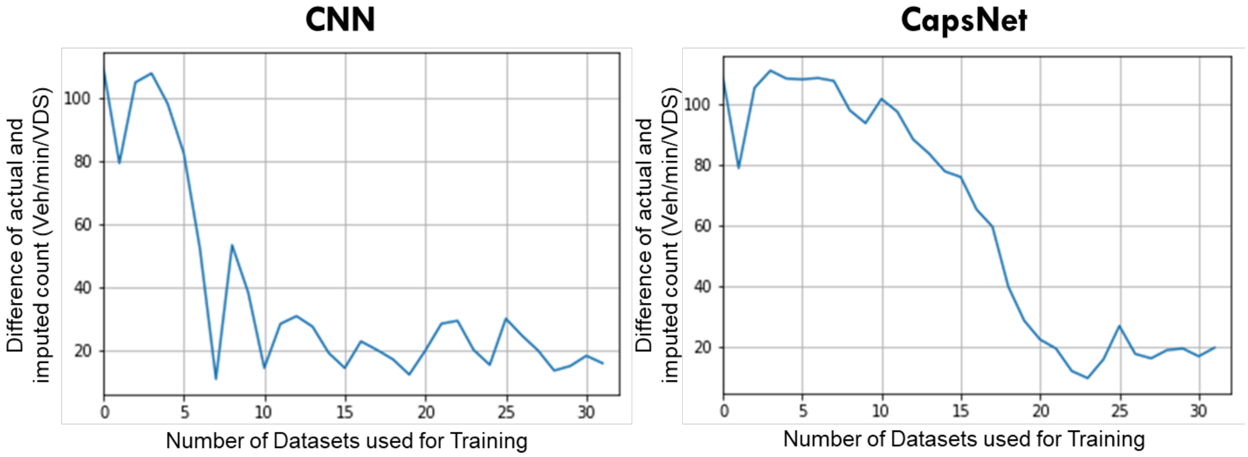


Figure 3-20 Effect of datasets on training

Table 3-3 shows the results from the ML-based imputation. The error or difference is measured based on: (i) the actual data (available in the PeMS dataset), and (ii) the imputed data (imputed by CNN and

CapsNet). The findings are based on the evaluation of every minute of data for each VDS available in the 23 test days. Based on the findings, the CNN model outperforms the CapsNet model with a difference of 6 veh/min/VDS error. The standard deviation of the error for CNN is also less compared to the CapsNet. Although CapsNet can capture more features, further evaluation is needed to study how those additional features can help to reduce the imputation errors.

Table 3-3 Imputation model performance

Imputation Method	Difference of actual and imputed count (Veh/min/VDS)				
	Mean	Std. deviation	25th Percentile	50th Percentile	75th Percentile
CNN	14.8	13.5	5.4	11.3	20.1
CapsNet	20.8	14.9	9.9	18.5	28.2

4.3. PEMS SPEED CALCULATION

To calculate speed from single loop detectors available in the study area, the g-factor based speed estimation method (as discussed in Section 3.1.1) is used. These g-factors, which are basically the average length of vehicles crossing a detector, influence the final calculated speed. A set of g-factors are used for each lane to identify which factor gives an acceptable range of error for almost all single loops. For a one-hour time interval, where loop data are aggregated for each minute, Figure 3-21 shows the error of calculated speed (i.e., speed using g-factor) and actual speed (i.e., speed from the simulation) for a specific g-factor (i.e., 22 ft.) for all loops in lane 1 of the freeway mainlines. The error is acceptable if: (i) the 50th percentile value of the error range is close to 0, and (ii) the sample size is near-equal for the overestimated and underestimated values.

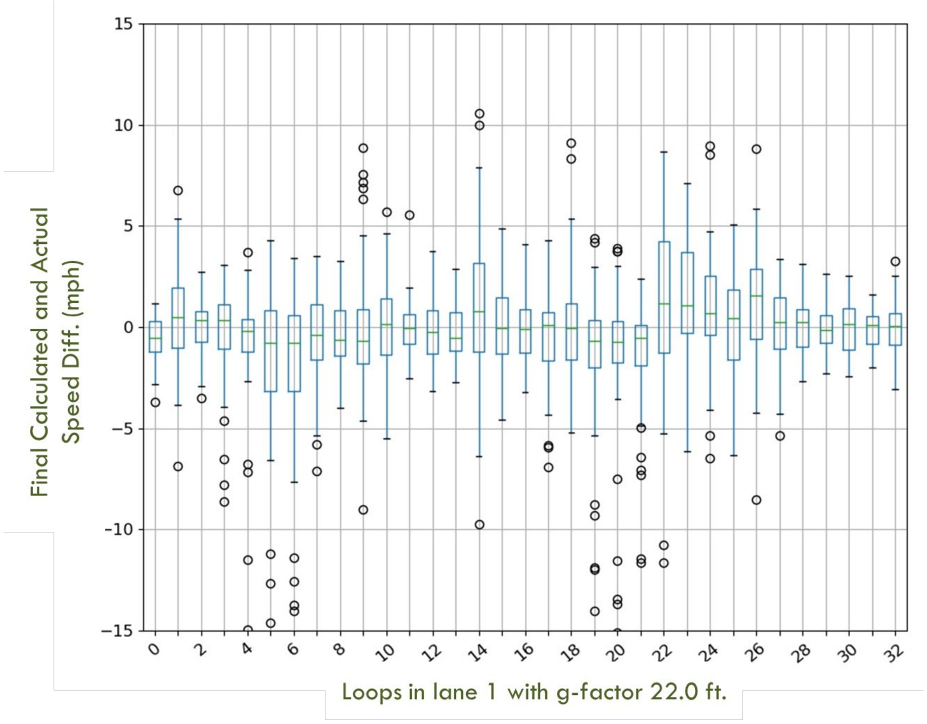


Figure 3-21 Calculated speed error for lane 1 in freeway mainline

A similar analysis is conducted for single loops in both freeway mainline and freeway-freeway connectors. For the freeway mainline, the final g-factor values for six lanes are found to be 22, 22, 26, 25, 24, and 23 ft., from the leftmost to the rightmost lane. For freeway-freeway connectors, the left and right lane g-factor values are both 23 ft.

Table 3-4 shows the findings for the absolute difference between g-factor estimated speed and the simulated speed for the morning congested scenario. Here data are aggregated for all single loops in the freeway mainline. For the congested scenario, the mean speed difference is 2.32 mph with a standard deviation of 2.89 mph, which means that in simulation, the g-factor method can generate speed values that are very close to the simulated speeds.

Table 3-4 PeMS speed estimation result

Simulation Scenario	Absolute difference of final calculated and actual speed (mph)					
	Sample	Mean	Std. deviation	25th Percentile	50th Percentile	75th Percentile
Morning congestion	8496	2.32	2.89	0.68	1.49	2.85

4.4. FREEWAY MAINLINE PERFORMANCE MEASURES

This section discusses the findings of both PeMS and hybrid methods to calculate the performance measures.

4.4.1. CONFLATION FOR THE HYBRID METHOD

Data conflation projects data (from different sources) onto the same spatial reference system. Figure 3-22 shows a sample representation of the flow data availability from VDS along the westbound portion of I-210. These flow data are conflated to the desired cells with a length of 0.25 miles. After applying the flow conflation using GASM and C-GASM methods, conflated flow at these cells is available. The error of the flow conflation method can be calculated based on the actual flow data available from the loops which are placed at those cell locations.

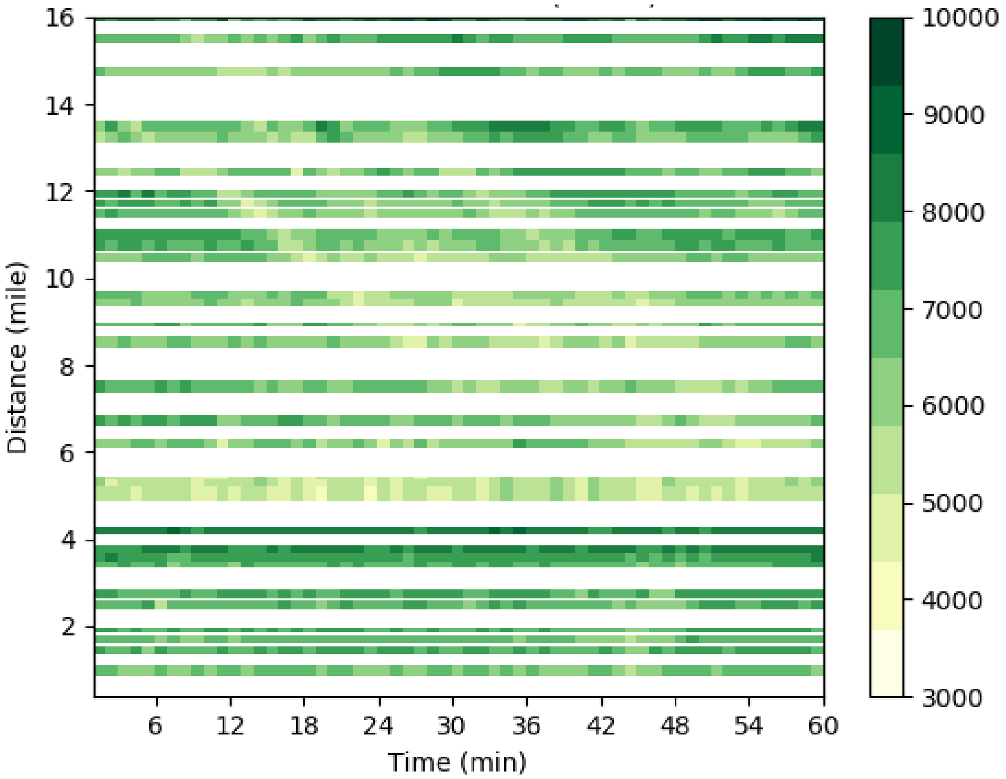


Figure 3-22 Sample available data from VDS along I-210 westbound

The v_{cong} and v_{ff} values are 80 and -25 kmph. The spatial and temporal smoothing widths, δ and μ, are found based on trial and error. Table 4-3 shows the mean absolute error and mean absolute percentage error for every minute at all cell locations for the morning peak period. In the analysis, the error values are the average error of two simulated replications. As C-GASM exhibited superior performance, it is used in the final application of performance measure estimation.

Table 3-5 Error of flow conflation at morning peak

Conflation Method	Mean Absolute Error (veh/hr)	Mean Absolute Percentage Error (%)
C-GASM	482	8
GASM	515	8.2

4.4.2. COMPARISON OF PEMS AND HYBRID METHODS

Following Section 3.1.2, performance measures are estimated using the PeMS method. In the hybrid method, once both flow and travel time data are conflated, the performance measures (VMT, VHT, and VHD) are calculated at the desired cell locations. Finally, all values along the whole freeway are summed up to get the final VMT, VHT, and VHD for the whole freeway.

Table 3-6 shows the VMT, VHT, and VHD values for all scenarios, and methods. Simulated ground truth data is calculated from the model. The simulation provides space-mean speed and vehicle count data for the simulated sections, which are used to get the base VMT, VHT, and VHD. For each scenario, VMT calculated with the hybrid method is closer to the simulation ground truth, than that calculated with the PeMS method. A small VHD is observed during the night off-peak scenario as an artifact of the cell-based travel time conflation, however, this can be considered as negligible.

Table 3-6 Performance measures for freeway mainline

Scenario	Calculation Method	VMT (veh-mile)	VHT (veh-hour)	VHD (veh-hour)
Before Morning Peak	SGT*	94598.56	2924.55	1374.44
	PeMS	92624.45	2788.03	1270.74
	Hybrid	93316.80	2918.89	1456.94
Morning Peak	SGT	80237.48	4689.71	3354.89
	PeMS	77405.69	4338.11	3051.96
	Hybrid	78128.70	4630.18	3366.20
Noon Time	SGT	93634.13	2598.23	1064.56
	PeMS	92035.59	2641.62	1119.39
	Hybrid	92675.72	2507.76	1054.24
Afternoon Peak	SGT	91021.54	3235.14	1743.29
	PeMS	89291.75	2975.35	1514.38
	Hybrid	90060.30	3098.24	1696.65
Night Off-peak	SGT	56544.19	822.31	0.04
	PeMS	55889.45	837.67	0.00
	Hybrid	56212.10	741.84	1.14

** SGT = simulated ground truth*

Figure 3-23 shows the benefit of including third party data for VMT, VHT, and VHD calculations. For the morning and afternoon peaks, the hybrid method yields an improvement of 9% and 10.4%, respectively, for VHD compared to the PeMS method. Due to C-GASM based conflation, VMT also improves when using the hybrid method. Both PeMS and hybrid methods underestimate all the performance measures in both scenarios.

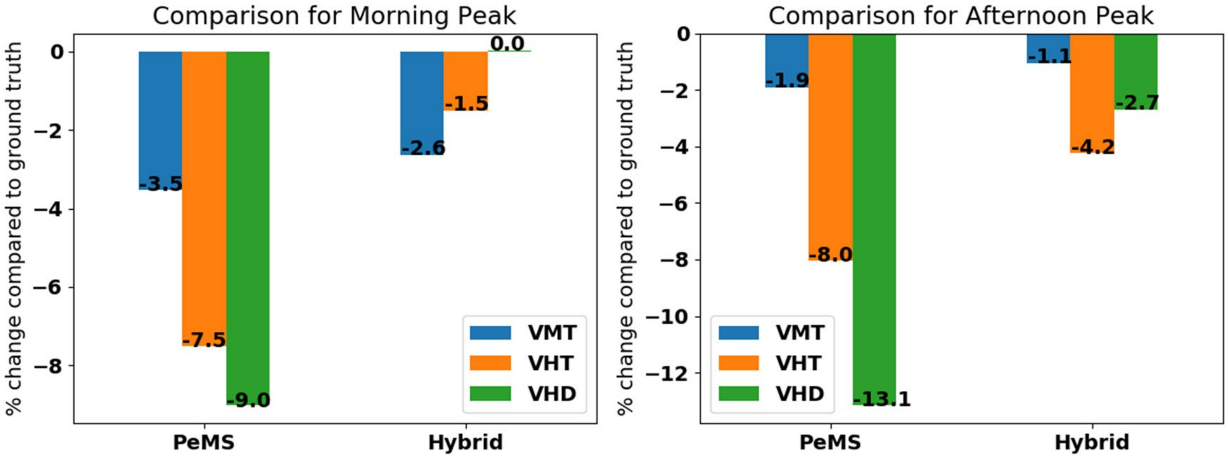


Figure 3-23 Peak period comparisons for freeway mainline

For the before-morning and noon scenarios, the hybrid method yields an improvement of 1.5% and 5.0%, respectively, for VHD compared to the PeMS method as shown in Figure 3-24. Due to C-GASM based conflation, VMT also improves when using the hybrid method. For the noon scenario, PeMS overestimates both VHT and VHD, whereas the hybrid method underestimates them.

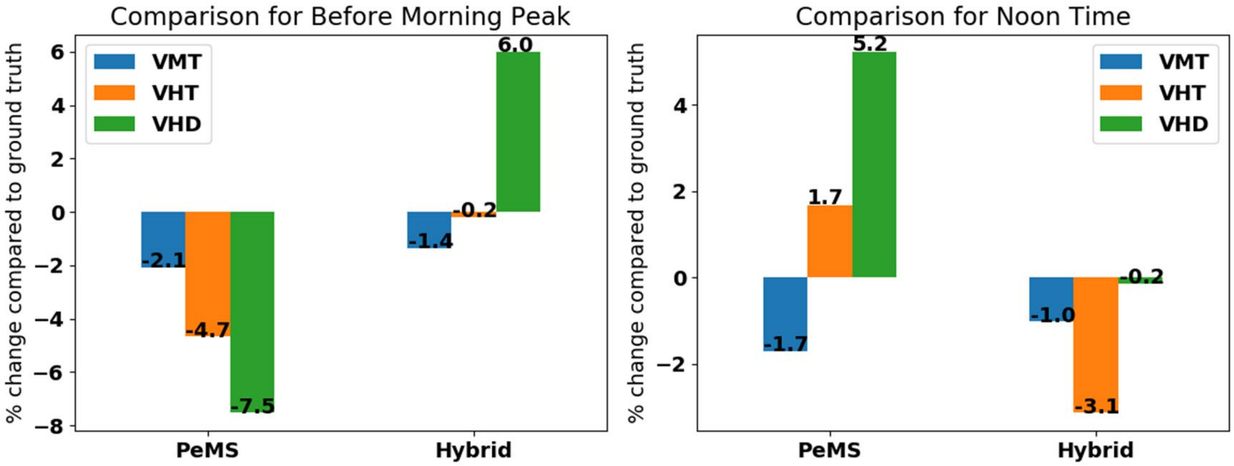


Figure 3-24 Other period comparisons for freeway mainline

4.5. FREEWAY-FREEWAY CONNECTOR PERFORMANCE MEASURES

Performance measures are calculated for the I-210 and I-605 freeway-freeway connectors using both the PeMS and hybrid methods.

Table 3-7 shows the findings for all the scenarios. Ground truth data is calculated from the simulation. The model provides space-mean speed and vehicle count data for the simulated sections, which are used to obtain ground truth VMT, VHT, and VHD. For almost every scenario, the inclusion of third-party travel time data improved the accuracy of the performance measures (VHT and VHD). The one exception is the afternoon peak. For connectors, no cells are used, and therefore VMT calculations for both hybrid and PeMS methods are identical, as the same data and calculation steps are used in both methods to compute VMT.

Table 3-7 Performance measures for freeway-freeway connector

Scenario	Calculation Method	VMT (veh-mile)	VHT (veh-hour)	VHD (veh-hour)
Before Morning Peak	SGT*	3929.53	113.78	43.15
	PeMS	3927.19	94.65	25.73
	Hybrid	3927.19	110.84	42.98
Morning Peak	SGT	4404.76	174.83	94.41
	PeMS	4364.82	124.32	50.44
	Hybrid	4364.82	167.69	90.82
Noon Off-peak	SGT	3657.01	68.69	6.37
	PeMS	3641.40	60.94	2.52
	Hybrid	3641.40	65.30	5.60
Afternoon Peak	SGT	4020.99	177.43	106.10
	PeMS	3870.43	167.75	101.28
	Hybrid	3870.44	166.81	101.55
Night Off-peak	SGT	3338.02	53.09	0.00
	PeMS	3328.17	48.67	0.00
	Hybrid	3328.17	49.89	0.00

** SGT = simulated ground truth*

During the morning peak, before the morning peak, and at noon scenarios, noticeable improvements can be achieved by using hybrid data. As shown in Figure 3-25, the VHT and VHD error reductions are 29.8% and 61%, respectively during the morning peak. During the afternoon, major congestion occurs in the Northbound I-605 to Eastbound I-210 connector in the bottom-right, as shown in Figure 3-26 in red. For this connector, point speeds captured by loops on the connector matches the space-mean speed data calculated using the third-party travel times.

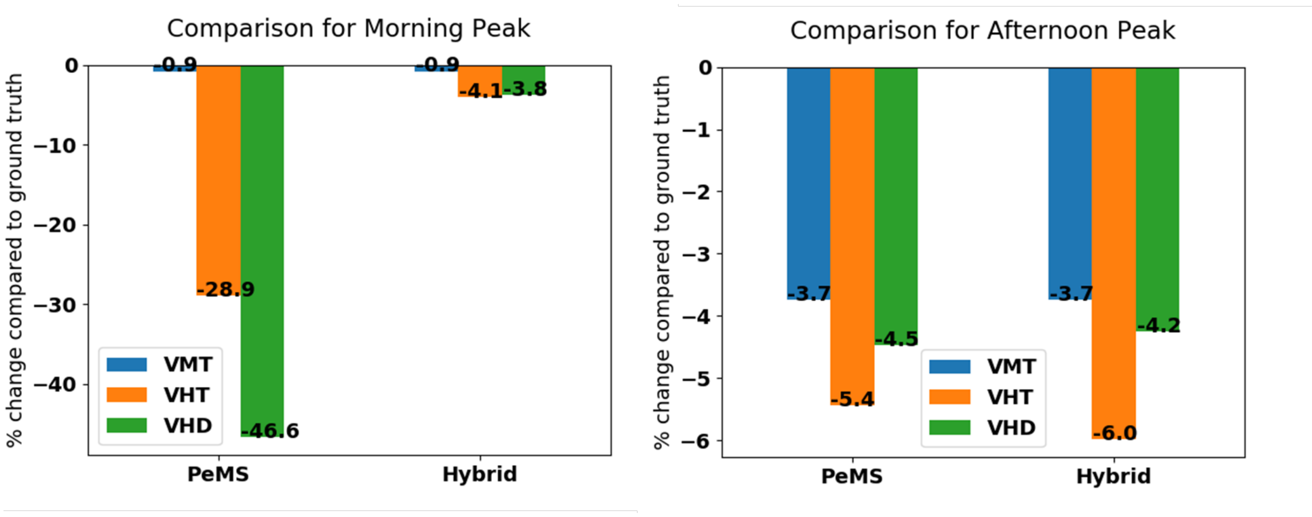


Figure 3-25 Peak period comparisons for freeway-freeway connectors

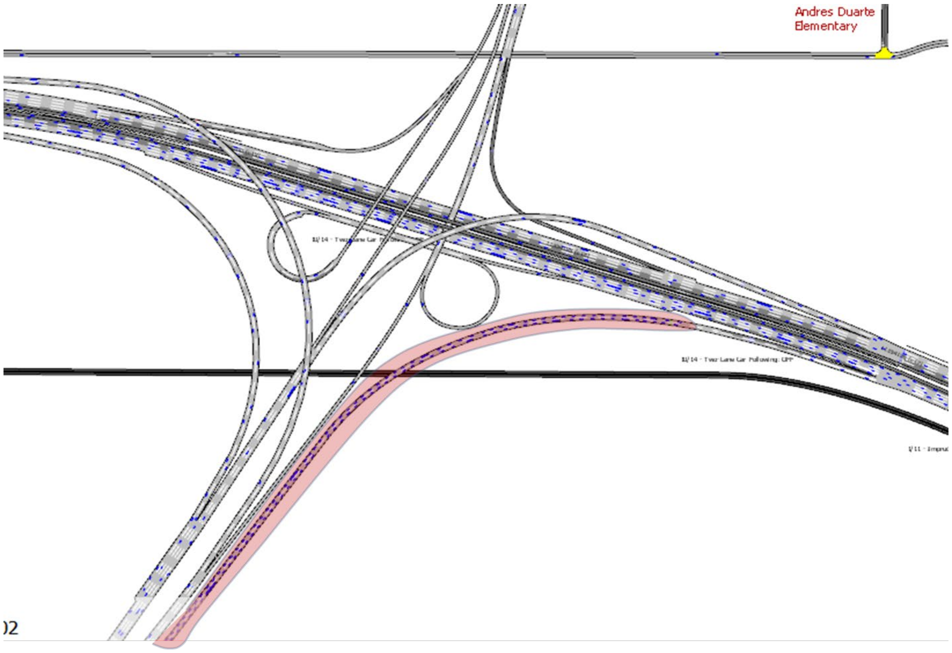


Figure 3-26 Freeway-freeway connector congestion during afternoon peak

Improvements are observed for both the before-morning peak and noon scenarios. As shown in Figure 3-27 for before-morning and noon, error reductions are 40% and 52%, respectively, if VHD is calculated with data from both VDS and third-party.

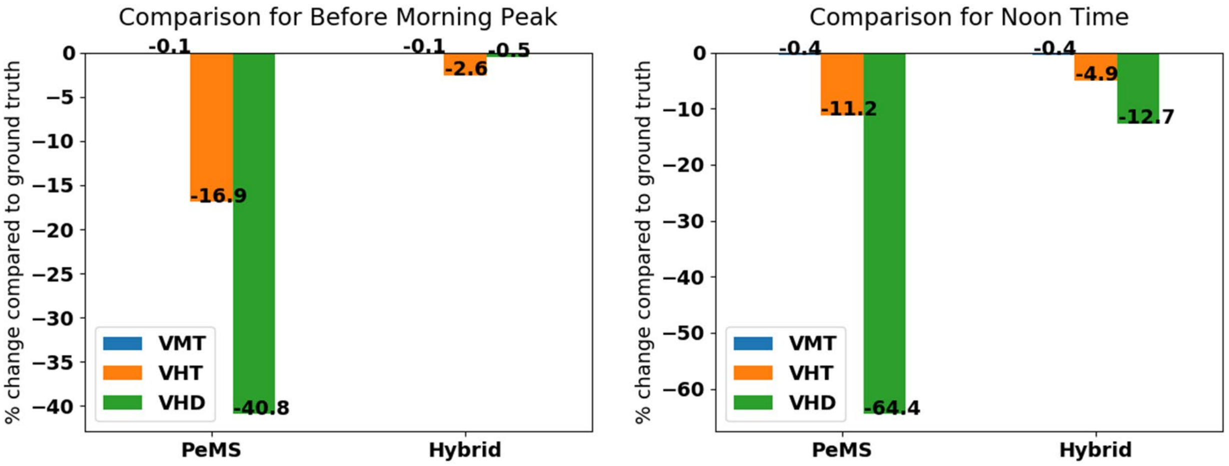


Figure 3-27 Other period comparisons for freeway-freeway connectors

4.6. CONCLUSIONS

The objectives of this research are to develop and evaluate a framework to calculate performance measures using a mix of data from multiple data sources. A literature review is conducted to study the current practices of data conflation and data fusion. The experimental design includes the use of the I-210 simulated freeway. Research is conducted for freeway mainline and freeway-freeway connectors using the simulated model. Both g-factor based point speed estimation (for both mainline and connectors) and limited probe data availability (for freeway mainlines) are considered to mimic real-world conditions.

The summary study findings are as follows:

1) The experiment conducted in this study for the ML-based imputation is performed for a specific scenario with 16 VDS having data and 9 VDS missing data. Based on the findings, these data-driven models have the potential to impute missing count data at such VDS locations. However, the development, validation, and implementation of general data-driven models to handle multiple scenarios require additional effort for further refinement.
2) Based on the experiment for morning congestion, the g-factor based speed calculation for single loops can generate a good estimation of point speed in simulation. However, using the point speed to calculate freeway wide VMT, VHT, and VHD can produce erroneous results, even if the point speed is measured with other sensors (e.g., dual loops).
3) For freeway mainlines, fusing data from third-party vendors helps to get a better estimation of the performance measures in almost all scenarios. For the off-peak period, when there is no noticeable fluctuation in demand, single point detector-based estimation is enough to estimate the performance measures. This happens as the point-speed at off-peak periods is very close to the space mean speed for the link.
4) For freeway-freeway connectors, similar results are observed, except that the improvements can be much greater. Many of the connectors are metered. The effect of metering cannot, in general, be captured by an upstream VDS on the connector. In cases where the traffic state at the VDS happened to be representative of that on the connector, then fusing data from third-party sources yielded no substantial improvement.

Chapter 4

Incorporating Third Party Data

1. OVERVIEW

This chapter evaluates methods for estimating VHD (Vehicle Hours of Delay) in multiple ways using a flexible mix of both traditional sensor data and third-party, probe-based mobile data. The following topics are discussed: Point-based detector deployment strategy, evaluating VHD estimation using hybrid data fusion along fully instrumented freeways; and opportunities for improved coverage using hybrid data along limited instrumented freeways.

1.1. APPROACH

It is often unfeasible for detector coverage to extend across an entire road network in an agency's jurisdiction. Due to resource constraints, less critical roadways have reduced or limited instrumentation coverage in the network. The purpose of this section is to understand the extent to which error is introduced while calculating VHD using third-party vendor data when point-based sensor data are either:

a) Selectively removed and compensated for with third-party data, or
b) Entirely supplemented with third-party data (e.g., roadways with limited instrumentation).

The extended coverage pertains to freeway segments only (e.g., remote rural highways) and not local arterials. The VHD calculation methods explored here may not be appropriate for local arterials since their traffic dynamics are strongly affected by traffic signals. Arterials may require different techniques as well as more detailed data in addition to commonly available travel times.

1.1.1. COMPENSATING FOR REDUCED POINT-SENSOR DATA

To evaluate the framework of point-based detector placement and demonstrate the potential ability for third-party data to compensate for the loss of VDS data, this report evaluates the effect of VDS removal both with and without third-party data. VDS are organized into FATVs. Each FATV contains two adjacent mainline VDS, one at its entrance, and one at its exit. Along a freeway, the VDS in each FATV are removed. VHD estimation accuracy is compared with and without third-party data and with and without VDS removed. Figure 4-1 illustrates the scenario of interest. Since a pair of VDSs are used to project traffic conditions along a segment, a pair of VDSs are removed at a time, repeating the comparison along the entire length of the corridor to generate a distribution of potential error.

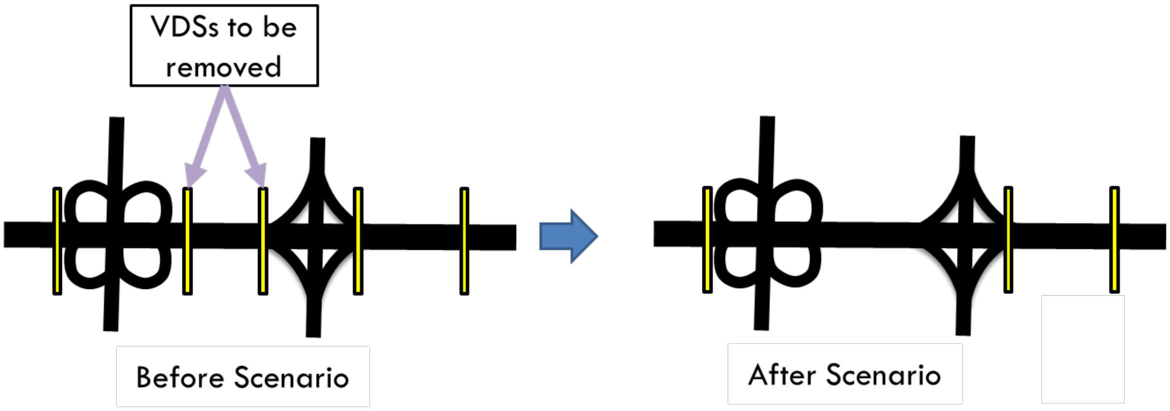


Figure 4-1 Point detector placement experiment

1.1.2. COVERAGE OF ROADWAYS WITH LIMITED INSTRUMENTATION

In addition to compensating for VDS loss or removal, another possible application of the third-party data is with roadways with limited instrumentation where no detectors are operating. Generally, the primary source of traffic count data on these corridors are aggregated AADT volume, which is typically estimated from sample counts from temporary sensors. To obtain the count at specific times of the day, generic flow profiles of hourly count distributions are used. Figure 4-2 shows such an hourly count distribution (Roess, Prassas and McShane, 2011).

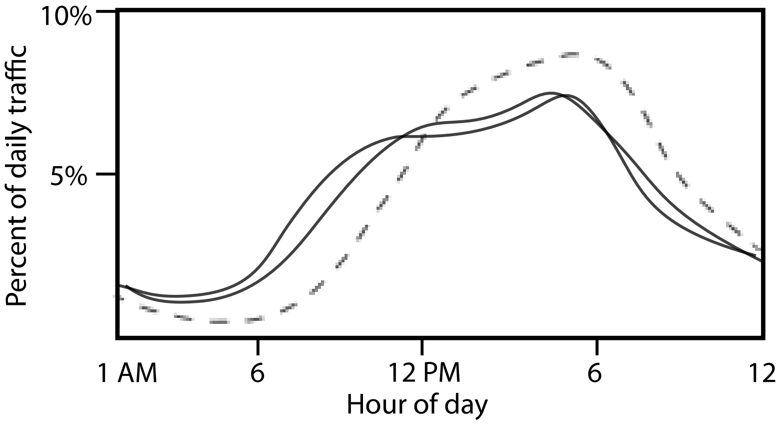


Figure 4-2 Sample hourly count distribution

These flow profiles can then be used to project AADT into hourly counts, which are then extrapolated into Vehicle Hours of Delay (VHD). The motivation of this research is to estimate the errors that can occur while estimating VHD using AADT values and third-party provided travel time.

1.1.3. SIMULATION STUDY

The proposed evaluation framework in this report was tested using data generated from a simulation model. This simulation of the I-210 corridor is described in Chapter 3, Section 3.3 Experimental Setup.

Simulated data allows the opportunity to observe data from all traffic along the entirety of the corridor. From this, researchers can extract select data subsets as if collected via point-based sensors or the spatially distributed third-party vendor data. This allows the natural introduction of error, such as with g-factor estimation, while still being able to compare against the original simulated "ground truth" (SGT) data.

VHD is calculated using the 65-mph speed threshold as the baseline. A more detailed discussion about the model, g-factor based speed estimation (for point detectors), and probe vehicle data simulation is discussed in the Chapter 3 Estimating Vehicle Hours of Delay. The absolute percentage error of both traditional and hybrid methods is calculated compared to SGT values which are obtained from the AIMSUN simulation. AIMSUN provides data for every link for any specified simulation time interval. The data includes vehicle count and space mean speed, which are used to calculate the SGT value for every simulation case.

1.2. SUMMARY FINDINGS

This technical report discusses a framework for two applications of third-party data: evaluating the use of third-party data to supplement VHD estimation with the removal of point-based detectors, and the ability to estimate delay on freeways with limited instrumentation. For the freeway mainline analysis, probe penetration of third-party data is assumed to be 5% and point detector speed is estimated from occupancy using a g-factor based method. Further discussions are included in the following sections to illustrate strategies to incorporate third-party vendor data.

Based on the analysis conducted in the research, there are two key findings:

- The hybrid method improved VHD estimates when FATV sensor groups were systematically removed (3.4% error with hybrid and 12.7% error with traditional data)
- Third-party data can estimate VHD using AADT on roadways with limited instrumentation, but the error is high (on average 41.3% error when using generic flow profiles, and 19.4% error with measured flow profiles).

The analysis is conducted for different times of day during weekday hours. Overall, in almost all the scenarios, findings show that including third-party data reduces VHD estimation error.

The findings suggest that including third-party data would be an improvement over traditional methods. There are two main ways this improvement could be used. The first would be to improve the accuracy of performance measures. Better measurement could enable better prioritization of resources and better investment decisions. Alternatively, the improvement could be used to gain a cost savings by using a different data mix while maintaining the current accuracy of performance measures.

2. EFFICIENT DEPLOYMENT OF POINT-BASED DETECTORS

This section first reviews the key advantages and disadvantages of conventional, point-based sensor systems. This discussion includes existing data pipelines, PeMS sensor location and meta-data configurations, and general point-based detector limitations. This is followed by a proposed framework and simulation study for evaluating the efficient deployment of point-based sensors and incorporating third-party vendor data. This section concludes with an analysis and discussion of the results of the new hybrid data strategy.

2.1. KEY VALUE OF POINT-BASED DETECTORS

Point-based sensors represent the backbone of traffic monitoring systems, collecting vital data at points across the network. Point detectors are widely deployed in California to collect traffic data but maintaining a vast network of point detectors can become costly and burdensome. With the exciting new potential of third-party vendor data, it is easy to imagine the obsoletion of point-based sensors. However, this is far from reality as point-based sensors still provide vital traffic data. Point-based sensors offer key advantages over emerging data sources:

- **Full counts** – Point-based detectors count every vehicle that passes (assuming negligible error), unlike mobile data which relies on a small sample of mobile-equipped probe vehicles. These full counts are necessary for travel demand and traffic census efforts, as well as traffic control operations, such as on/off-ramp monitoring, traveler information signs, and variable toll pricing. Mobile data samples can also introduce sampling and data-quality bias (e.g., commercial vs. non-commercial)
- **Precision** – Point-based sensors are physically placed, enabling consistent measurement at a specific location or across individual lanes. The precision afforded by point sensors is vital to traffic control and lane management strategies, such as HOV and HOT lane management.

2.2. KEY CHALLENGES OF POINT-BASED DETECTORS

Despite their advantages, point-based sensors possess an array of issues and limitations. Understanding the limitations of the point detectors will help to understand the sources and extent of error in the VHD calculation. These issues can be divided into four basic categories:

- **Fundamental limitations about collected data** – Point-based sensors typically only capture flow and occupancy, requiring speed and travel time to be estimated using assumed values (e.g., g-factor estimation) that introduce errors. The traffic between point-based sensor locations must be inferred, which if spaced too far apart can fail to accurately capture variable traffic flows, such as backward propagating waves.
- **Location discrepancies** – In PeMS, sensor locations are mapped to the linear postmile position of the central control boxes and communications equipment, and not the actual location of the sensors themselves. This creates challenges for analysis and modeling efforts.
- **Organizational structure** – The existing PeMS meta-data convention indicating sensor location on the freeway is inadequate for more complex urban freeway interchanges and is prone to

misconfiguration. This makes it difficult to check data consistency and data quality, to use data for modeling, and to incorporate third-party data.

- **Cost** – Point-based sensors are physical assets with costs associated with installation, maintenance, and operation. Sensors can malfunction due to any number of reasons, such as power outages, hardware communications failures, or general wear. These costs make the management of a large system difficult and cost prohibitive to expand network coverage to less utilized roadways, such as remote rural highways.

2.3. FRAMEWORK FOR EFFICIENT DEPLOYMENT

If a sensor is removed from a freeway, then traffic conditions must be inferred using data from remaining sensors further upstream and downstream. This potentially obscures traffic conditions in the vicinity of the removed sensor, resulting in erroneous VHD calculations due to traffic flow variations, such as backward propagating shockwaves or roadway geometry changes (i.e., decrease or increase in lanes). To measure this error, count locations are systematically removed and evaluated in a simulation study.

Considering the FATV sensor groups on I-210, VDS placement efficiency can be analyzed by removing a FATV sensor group and evaluating the resulting error incurred from the removal. Along the I-210, the subsets of FATVs can be thought of as two consecutive VDSs, where the first one represents the in-set and the second one represents the out-set. Figure 4-3 shows such arrangements to evaluate the efficient VDS placement framework.

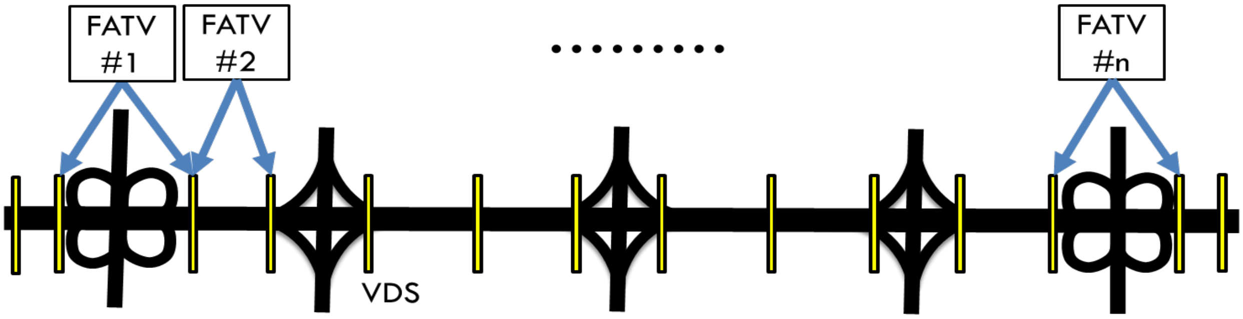


Figure 4-3 FATVs along I-210

In the first evaluation scenario, FATV #1 is removed and the other FATV locations remain to calculate VHD. Error is then calculated between the VHD estimate with all FATV locations and the VHD with one removed. In a subsequent scenario, FATV #2 is removed instead of FATV #1, calculating VHD using all remaining locations. This process is systematically continued across all FATV locations, effectively calculating VHD given the loss of each FATV location. The one caveat is that the very first and last VDSs are not removed in any evaluation scenario. These boundary VDS are needed for both flow and travel time conflation at the edge locations of the corridor.

To evaluate the potential impact of third-party data, VHD is estimated using two different methods, shown in Figure 4-4:

- **Traditional data method** – Data from only point detectors are used to estimate VHD.

- **Hybrid Data Method** – Data from a third-party is used with remaining point detector data.

In the traditional method, the upstream and downstream detectors become responsible for longer segments. In Figure 4-4, the traffic in the green shaded area on the left will be reorganized into the two extended road segments illustrated on the right.

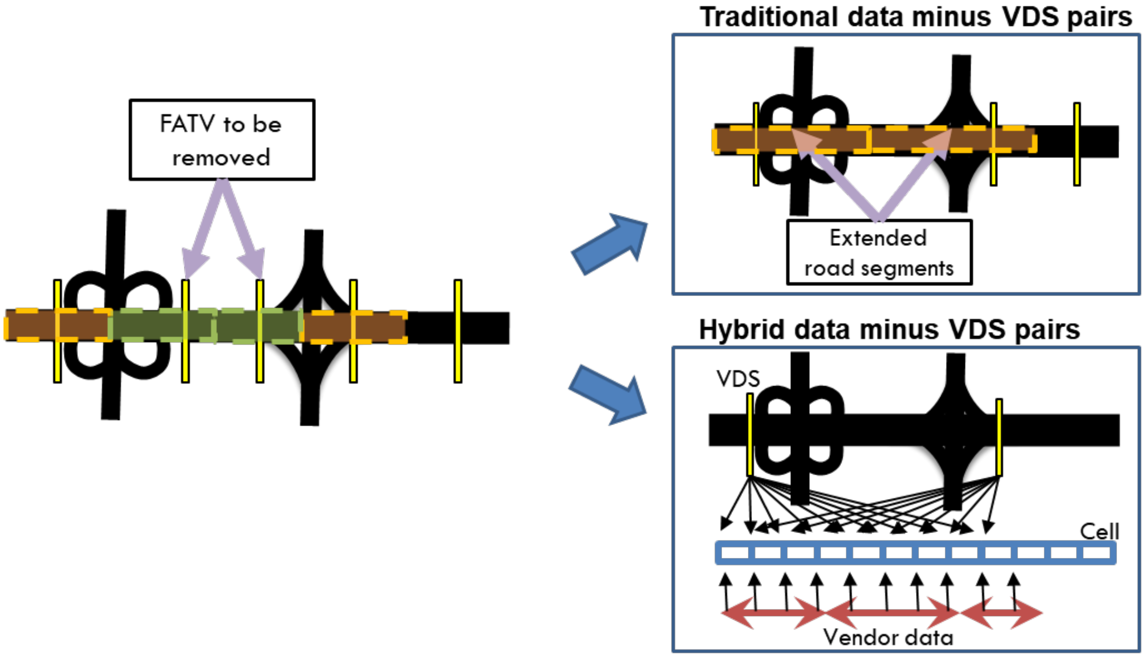


Figure 4-4 Framework for evaluating the efficient deployment of point-based sensors

In the hybrid method, data from both point detectors and third-party vendors are used to estimate VHD. As with the traditional data method, traffic information in the hybrid method is projected from further upstream and downstream but is also conflated onto the cells for third-party data fusion. The VHD calculation method (including flow conflation and data fusion) is described in Chapter 3: Estimating Vehicle Hours of Delay.

2.4. ANALYSIS AND DISCUSSION

The findings suggest that including third-party data would be an improvement over traditional methods. There are two main ways this improvement could be used. The first would be to improve the accuracy of performance measures. Better measurement could enable better prioritization of resources and better investment decisions. Alternatively, the improvement could be used to gain a cost savings by using a different data mix while maintaining the current accuracy of performance measures.

The initial analysis is conducted for the point detectors using the traditional method. Figure 4-5 shows the distribution of absolute percentage error for four different scenarios using the traditional and hybrid methods compared to the ground truth values. Using the traditional method for both peak periods (morning and afternoon), the mean absolute percentage errors are 14.6% and 18.8%; with the maximum error of 20.7% and 35.3%, respectively. Using the hybrid method for both peak periods (morning and afternoon) the mean absolute percentage errors are 0.9% and 4.6%; with the maximum error of 2.1% and 6.5%, respectively.

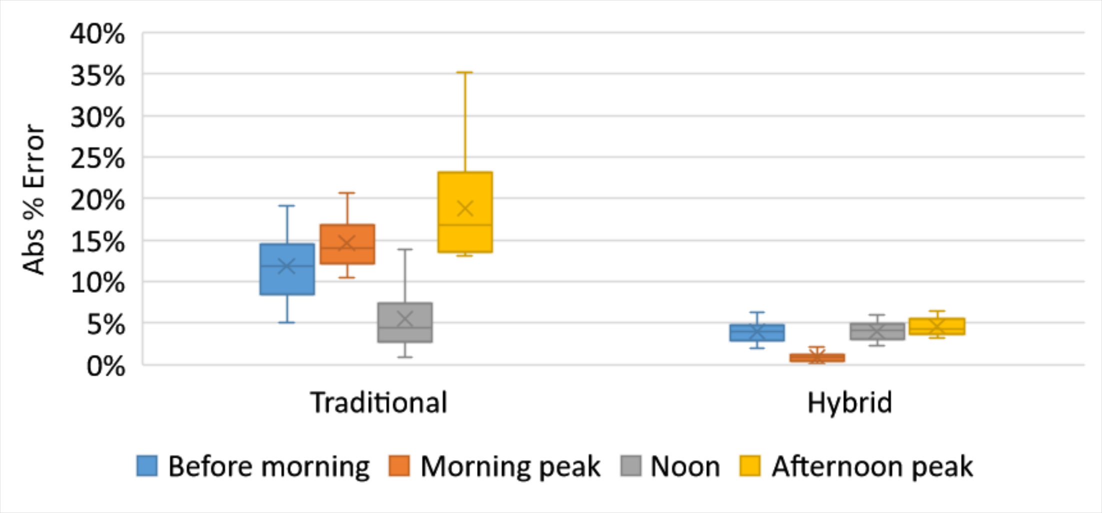


Figure 4-5 Abs percentage error for traditional and hybrid methods

The comparison between the traditional and hybrid methods in Figure 4-5 shows that the incorporation of third-party vendor-provided data can limit the absolute percentage error to a maximum of 6.5% in all scenarios. Whereas in the traditional method, the error can be as high as 35%. Figure 4-6 shows the distribution of absolute percentage errors of both methods for all scenarios. It is assumed that the errors follow a normal distribution. The mean absolute percentage error for the hybrid method is 3.4% while it is 12.7% for the traditional method. It is evident from the analysis that, although the FATVs are removed, the incorporation of third-party vendor-provided data can keep the error within a much lower range (mean % error less than 5%).

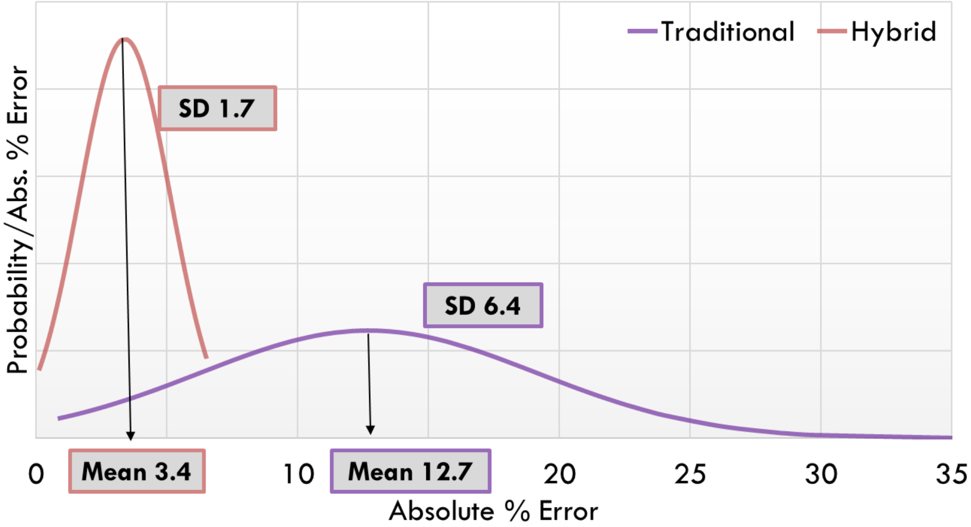


Figure 4-6 Abs percentage error distribution for FATV removals

The addition of third-party data yielded promising results. While third-party data did not achieve perfect results, it did effectively compensate for VDS loss, reducing the mean error by nearly four-fold from 12.7% to 3.4%. Whether this level of error is within acceptable tolerance limits depends on the specific application, but the improvement is substantial. All of this implies that in the context of delay estimation, third-party data can be used effectively to supplement a reduction in point sensors such as loops.

In general, the removal of FATV locations resulted in an increase in error as expected. What is not yet clear are the impacts from removing specific sensors. Some sensors provide a greater contribution to accuracy, perhaps near interchanges or along curves, and others result in less error if removed. It is challenging to generalize the results. They depend on the traffic patterns and congestion, in addition to the physical features of the road. However, we have provided a framework to evaluate the importance of sensor groups in both traditional and hybrid contexts. A summary of individual sensor error at Individual VDS Locations is included in Appendix A – VHD Calculation Error at Individual VDS Locations.

3. OPPORTUNITIES FOR IMPROVED COVERAGE

Due to resource constraints, it is often unfeasible to extend sensor coverage across an entire road network in an agency's jurisdiction. It is possible that third-party vendor data can cost-effectively extend coverage to freeways with limited instrumentation. The purpose of this section is to understand the extent of error when estimating VHD using third-party data along roadways with limited instrumentation.

The extended coverage pertains to freeway segments only (e.g., remote rural highways) and not local arterials. The VHD calculation methods explored here may not be appropriate for local arterials since their traffic dynamics are strongly affected by traffic signals. Arterials may require different techniques as well as more detailed data in addition to commonly available travel times.

3.1. EXTENSION OF COVERAGE TO FREEWAYS WITH LIMITED INSTRUMENTATION

Using third-party data, it is possible to expand the VHD estimation to freeways with limited instrumentation where only AADT data is available. This can be achieved by integrating the hourly segment travel times, obtained from third-party data, against the hourly traffic volumes. The hourly traffic volumes can be estimated as the hourly proportion of daily traffic (i.e., AADT) calculated from an hourly traffic flow profile. The traffic flow profile can either be a generic flow profile (e.g., from a traffic handbook) or measured from an available sensor nearby.

Since the flow profile is not the actual flow profile of the roadway in question, an error will be introduced. To determine the error incurred from VHD estimation on limited instrumentation roadways, a simulated approach similar to that described in Section 1.1.3 is used. The simulation generates the third-party data and simulated ground truth (SGT). VHD is then calculated using third-party travel time data, available AADT data, and a flow profile. Error is calculated between the simulated ground truth (SGT) and the estimated VHD. To determine the extent of error that might occur depending on flow profile sources, two test cases are explored:

- **Generic flow profile (Case 1)** – A generic flow profile is assumed to represent traffic flow along the entire segment.
- **Measured flow profile (Case 2)** – A measured flow profile from a nearby VDS is assumed to represent traffic along the entire segment.

In both cases, AADT and third-party data are available from somewhere along the segment and a flow profile is assumed to represent traffic flow for the entire segment. The key difference is whether the flow profile is generic, such as from a traffic handbook, or if it is a measured flow profile from a nearby sensor.

Figure 4-7 shows the generic flow distribution used in Case 1. The time-of-day simulation points (before morning, morning peak, noon, and afternoon peak) where VHD is calculated are highlighted in the figure.

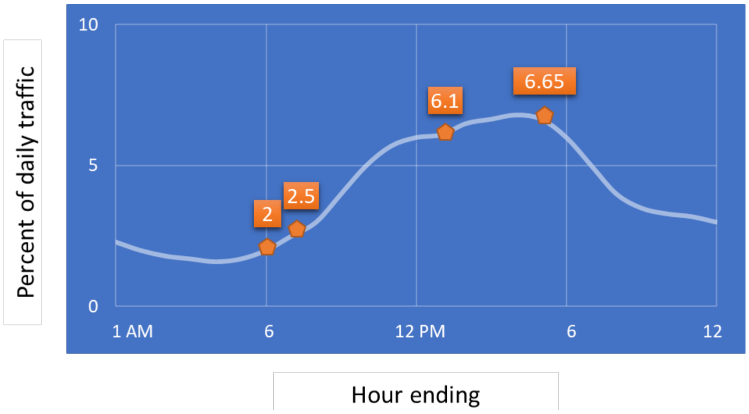


Figure 4-7 Generic hourly flow distribution

To calculate a range of errors, several AADT values from different points along the I-210 freeway are used to estimate delay, where AADT values are from real-world data. The AADT data was collected on August 1st, 2019 from the PeMS database, shown in Table 4-1 (Caltrans, 2020a).

Table 4-1 AADT values considered for Case 1

VDS ID	AADT
764137	117,446
717637	103,593
717644	100,789
717653	101,769
717669	91,726
761342	87,442
718210	86,933
769702	73,981
769722	78,366
717673	74,989

In Case 2, the measured flow profiles are based on data from nearby VDS locations within 10 miles upstream and downstream from the I-210 study corridor, but not on the corridor itself. This accounts for error varying depending on sensor proximity to the corridor. Once again using real-world data, the flow profile VDS locations, listed in Table 4-2, were collected on September 16th, 2020.

Table 4-2 VDS ID for Case 2

Serial Number	VDS ID
1	770200
2	770141
3	769867
4	770386
5	717694
6	769953
7	718047
8	769136
9	767940
10	768000

Figure 4-9 and Figure 4-9 show the flow profiles for the 10 nearby VDS locations, demonstrating the range of flow profile variation upstream and downstream from the study corridor.

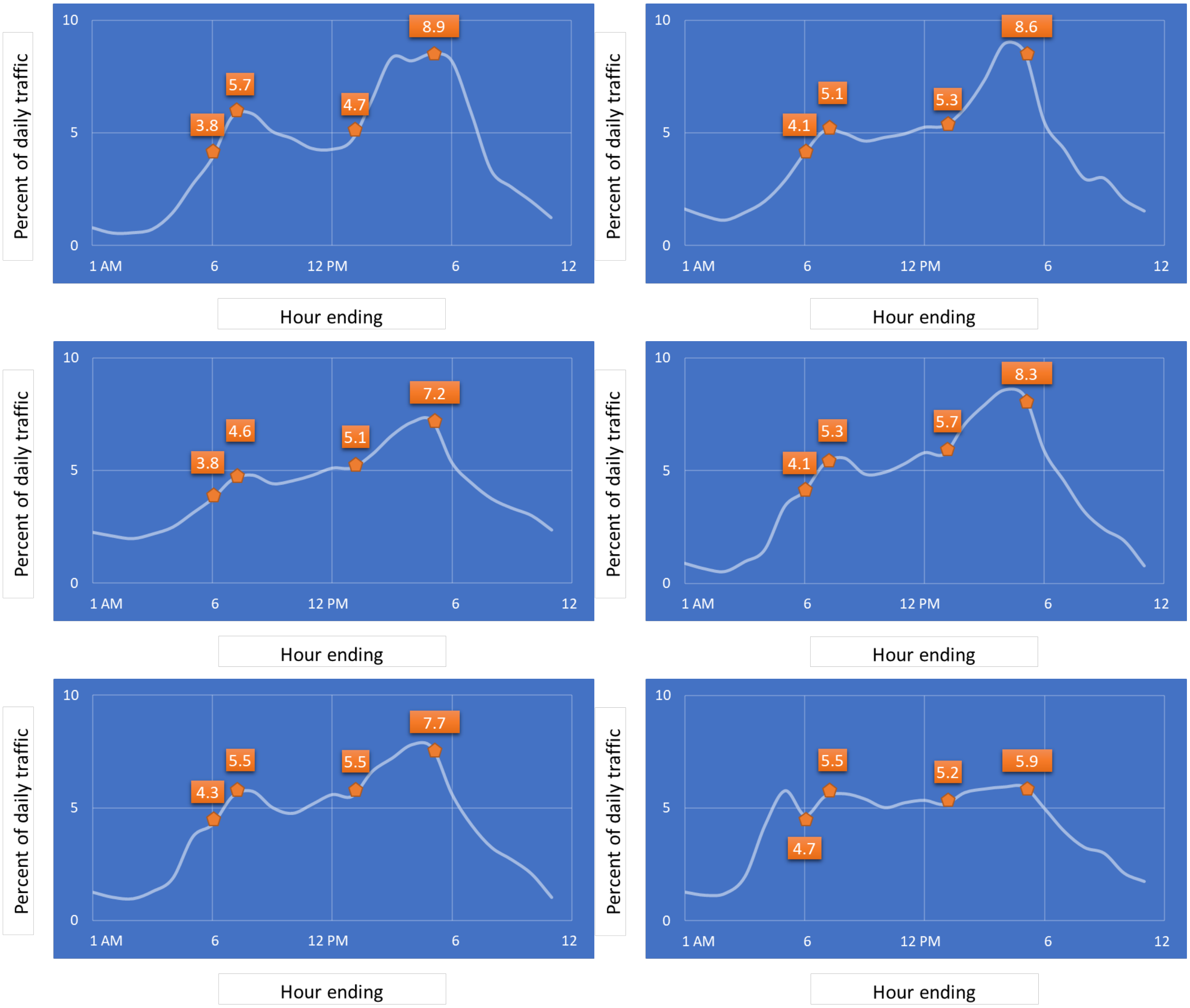


Figure 4-8 Measured flow profiles from I-210 up and downstream VDS

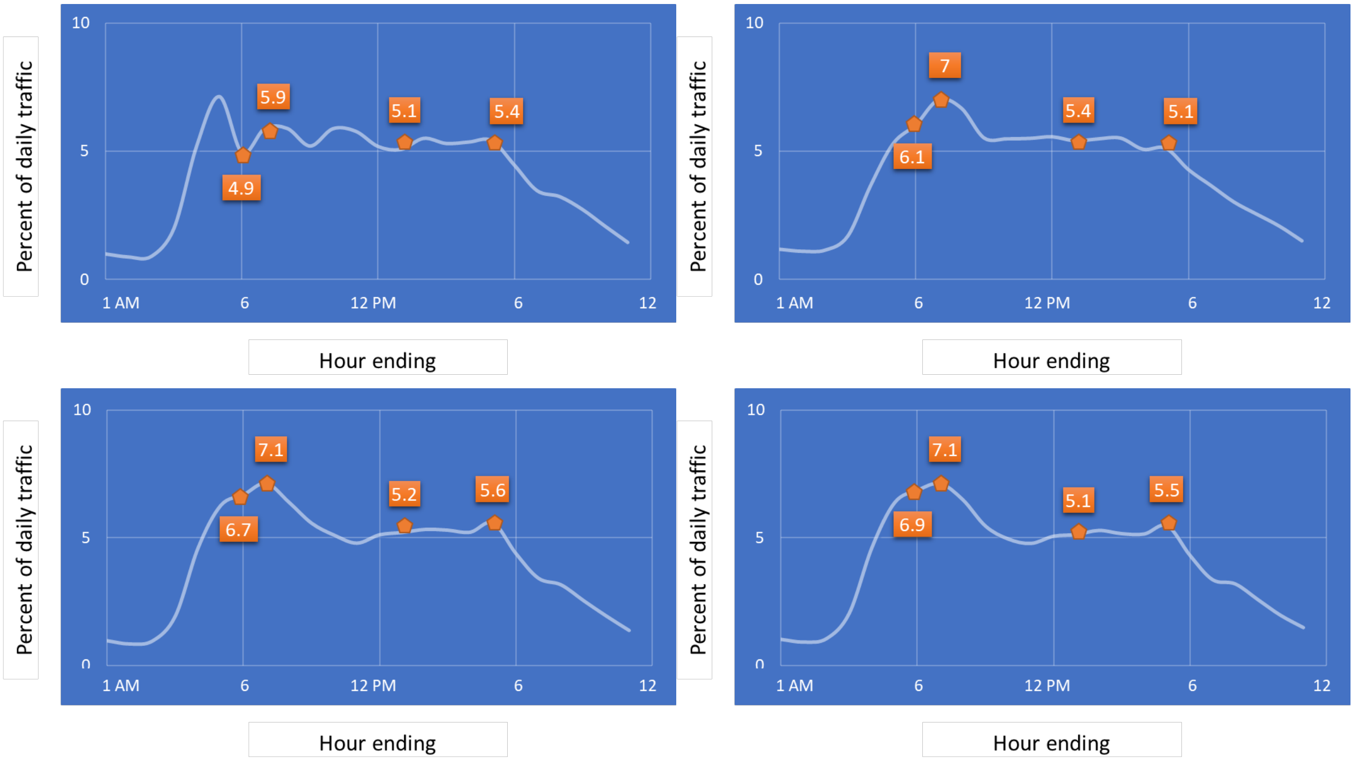


Figure 4-9 Measured flow profiles from I-210 up and downstream VDS (continued)

3.2. ANALYSIS AND DISCUSSION

Figure 4-10 shows the range of errors for four different scenarios compared between the two cases. For Case 1, the means of absolute percentage error are 72%, 60%, 19%, and 14% for the before morning peak, morning peak, noon time, and afternoon peak, respectively. For Case 2, the means of absolute percentage error are 22%, 23%, 10%, and 23% for the same times. In both cases, the errors are relatively high compared to the scenario where point-based sensor data are available for the corridor. However, the errors in Case 1 with the generic flow profile were substantially higher in the morning hours. This presents a major weakness to a simple generic profile introducing a large source of error.

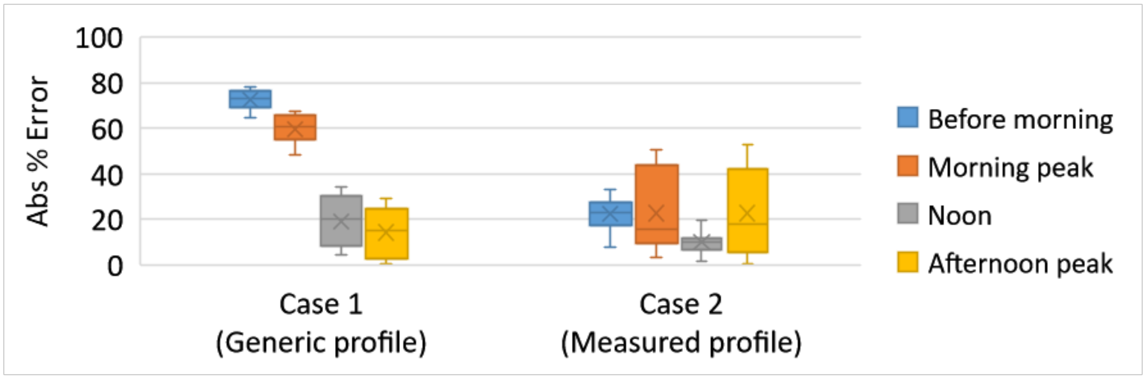


Figure 4-10 Abs percentage error with a flow profile

For an overall comparison of approaches, Figure 4-11 shows the absolute percentage error when using traditional point-based sensor data only with removed FATVs, hybrid data with removed FATVs, and for the two limited instrumentation cases. It should be noted that while most distributions appear to follow a Gaussian distribution, this cannot be confirmed for the limited instrumentation Case 2. The large gaps in data points, particularly in the center, make it difficult to confirm. Regardless, the distribution still demonstrates the relative mean and spread of error.

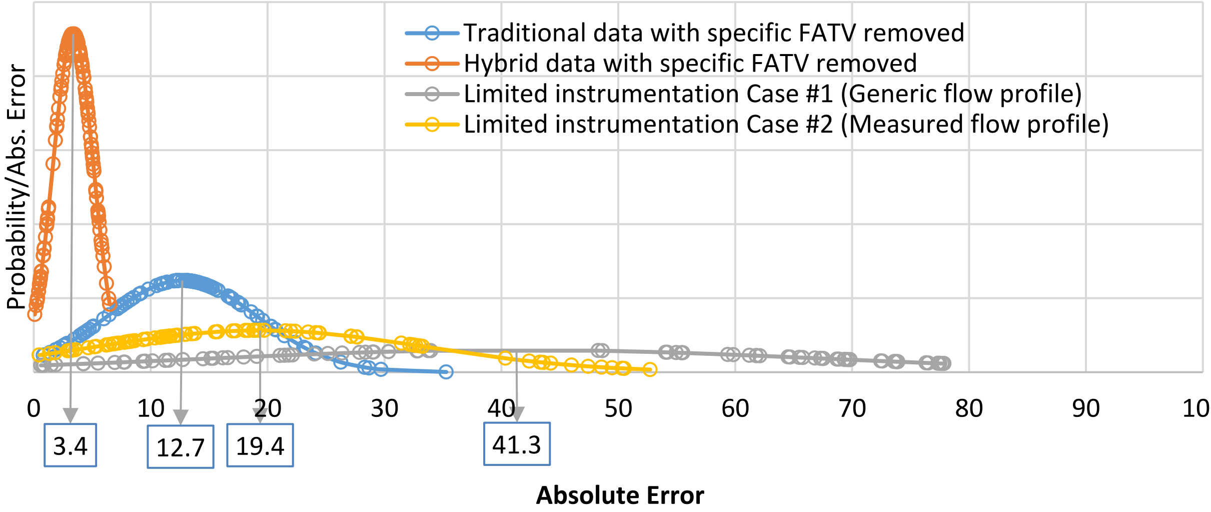


Figure 4-11 Abs percentage error distribution using different data sources

Overall, the hybrid data case yielded the best level of error at only 3.4% compared to 12.7%, as discussed previously in Section 2.4 of this chapter. However, both levels of error are relatively low compared to cases with limited instrumentation. The latter yield means of absolute error of 19.4% and 41.3% for the measured and generic profile cases, respectively. However, if only rough estimates of delay are required for some roadways, such high levels of error might be acceptable.

Chapter 5

Strategy to Incorporate Third-Party Data

1. OVERVIEW

In this chapter, proposed strategies for incorporating third-party data are discussed. First is a review and comparison of third-party and point-based data. This is followed by proposed strategies for a hybrid data framework, and a third-party data roadmap strategy.

Accurate and precise traffic information is vital to the operation, planning, and governance of roadway networks. Traffic information is used for a range of purposes, from high level decision making and design, down to providing traffic information to commuters. In recent years as capacity expansion has begun reaching practical limitations, the focus has shifted away from capacity expansion and towards traffic management strategies, making the need for high quality traffic data paramount.

Third-party data (from mobile devices or other sources) could be incorporated into transportation management systems to complement data currently collected by Caltrans. Employed sensibly, such a hybrid data strategy could augment information availability and quality with potential cost savings compared to current business practices. Increasing information access would improve Caltrans' and its local partners' ability to manage roadway traffic in a variety of existing and emerging ways to balance demand and flow to better utilize infrastructure investments.

1.1. DATA COMPARISON

Point-based sensors, such as induction loops, have been the backbone of traffic monitoring systems for decades and provide the basic function of counting vehicles that pass a point over a given time duration. However, point-sensors are ultimately limited by only collecting data from a singular point, and not continuously along a roadway segment. Data captured at a specific point means that traffic conditions upstream and downstream from the sensor must be extrapolated or projected. This conflation introduces potential sources of error if traffic conditions are not consistent between sensors. This is particularly true when calculating travel time if, for example, a slow-down or some backward propagating wave occurs causing traffic flow conditions to be non-uniform along the segment. Of course, accuracy can be improved by increasing sensor frequency and decreasing the distance between sensor locations, but this can quickly become cost prohibitive. Another critical source of estimation error is that single-loop sensors cannot estimate speed directly. Instead, speed is estimated using some assumed average vehicle length and occupancy to calculate speed (e.g., g-factor estimation). This assumption introduces error because vehicle length varies, which is further compounded by any occupancy measurement error.

Rather than calculating the speed over a span of time at a single point, the mean speed spanning a length of the roadway can be calculated from multiple vehicles simultaneously. This yields a more reliable speed calculation, and thus travel time. Prior to enough market penetration of location-tracked mobile devices (e.g., GPS enabled smartphones or connected vehicles), data for such a calculation were difficult to capture. Now with the proliferation of mobile devices, third-party vendor data is readily available, offering accurate speed and travel time estimates. However, third-party mobile data generally relies on a small sample of mobile-equipped probe vehicles within the traffic stream, which makes third-party data less appropriate for traffic control applications and introduces potential sampling bias. For example, penetration rate may vary across income levels, across vehicle types such as commercial vs. non-commercial traffic, regions with poor cellular coverage, and due to location-sharing privacy concerns. Point-based sensors in contrast provide full traffic counts.

The inherent characteristics of the two different data sources, fixed point-based versus mobile detection, make the hybrid calculation a challenging task. The third-party provided data must be conflated on the same coordinate system to be fused with the VDS data. However, existing disparities between coordinate systems used by the vendors and transportation agencies can hinder the desired conflation. These differences can arise from a change in linear reference systems, freeway segmentation definitions, data coverage, and roadway geometries (Chen, 2019). Third-party data typically do not disclose the available probe vehicle penetration levels. For freeways with mainline and HOV lanes, third-party data will typically provide aggregated travel time information, which does not distinguish travel times between the mainline and HOV lanes. In contrast, point detectors, like VDS, provide lane-specific data and full counts, but only flow and occupancy values are available if only single loop detectors are used.

Overall, both data sources provide trade-off benefits and dis-benefits. Third-party mobile data offers advantages over point sensors for longer road segments, reducing the need for frequently spaced detector stations and potentially extending coverage to remote roadways with limited instrumentation. But point-sensors offer full counts and fixed precision (e.g., specific lanes). These advantages for fixed sensors are necessary for more complex traffic conditions or where traffic control is required, such as actuation for traffic signals located at the intersection of freeway ramps and arterial streets, ramp metering control at freeway on-ramps, and pricing for HOT lanes.

2. PROPOSED HYBRID DATA FRAMEWORK

This overall framework incorporates four steps as shown in Figure 5-1. At first, the data is acquired from both traditional point detectors (i.e., VDS) and third-party vendors. An initial data quality check is conducted to evaluate whether the data are usable to estimate performance measures. After performing the quality control on the data, both flow and travel time data are conflated to project them onto the desired cell. After having both flow and travel time data conflated, data fusion is performed to calculate the desired performance measures.

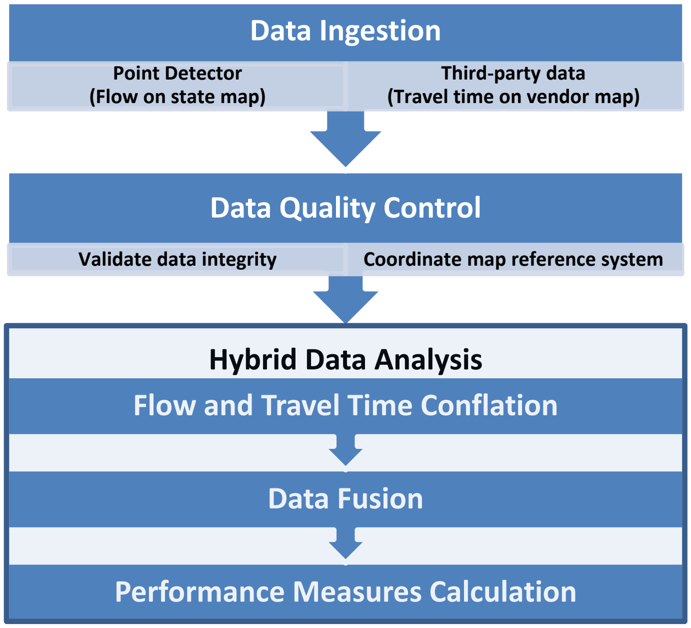


Figure 5-1 Steps for performance measurement estimation

The proposed framework contains three overall components, Data ingestion, Data Quality Control and Coordination, and the Hybrid Analysis itself. The following three subsections will discuss the proposed strategies associated with each step.

- **Proposed Organizational Approach for Data Quality Control and Coordination using FATV concept** – FATV is a proposed remedy to improve data quality and provide precise location information that is needed for incorporating third-party data and ensuring data integrity.

- **Proposed strategy for VHD in PeMS** – A proposed strategy for VHD in PeMS and the potential for improved point-based sensor data and coverage extension.

- **Third Party Data Roadmap** – Presents a strategy and a two-part plan for Caltrans to move forward with hybrid data implementation.

The three proposed strategies will be discussed in the following sections.

2.1. PROPOSED ORGANIZATIONAL APPROACH WITH FATV CONCEPT

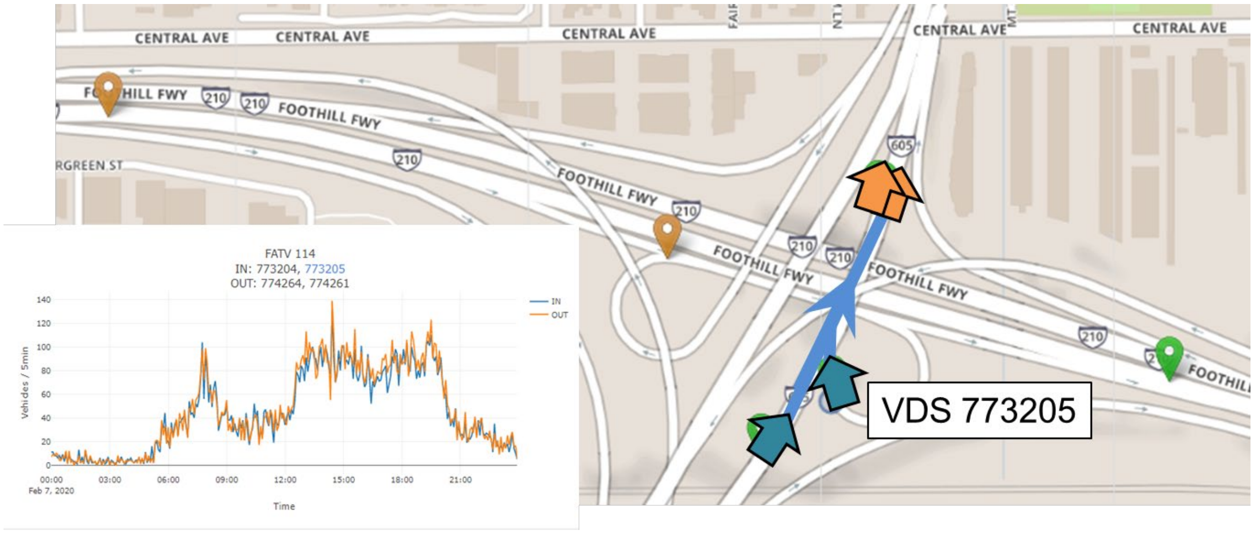
As previously discussed, the current organizational configuration of point-based sensor meta-data is prone to misconfiguration and fails to detect anomalies. This makes it difficult not only to utilize data for analysis and modeling, but also in going forward with third-party data integration. While fundamental limitations of point-based sensors are unavoidable, issues relating to location discrepancies and meta-data organizational structure can be mitigated.

Short of a complete overhaul of the organizational structure and geospatial locations of point-based sensors, a proposed remedy is the FATV concept. FATV is an organizational hierarchy with two primary objectives:

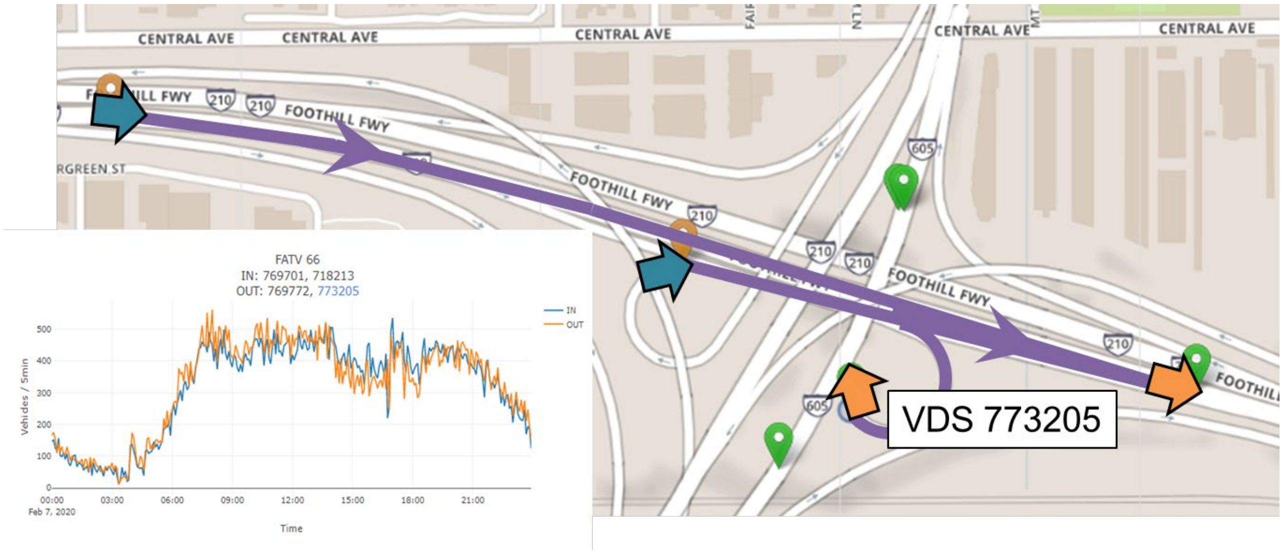
- To support automated validation of sensor data, and
- To spatially organize sensors for future fusion with third-party data.

The basic approach of FATV is to compare the ingress and egress vehicle counts of a group of detectors. That is, to compare the total number of vehicles entering and exiting a freeway segment. Each FATV set can be defined by the union of two subsets of point detectors, an in-set, and out-set where a vehicle that arrives from the in-set cannot exit the network without traveling through the out-set. For any FATV set, the vehicle total count at any given interval can be written as $C_{IN}+ C_{CN} = C_{OUT} + C_{RM}$. Here for a FATV confined area C_{CN} and C_{RM} are the numbers of vehicles contained at the beginning of the time interval, and remaining at the end of the interval, respectively. C_{IN} and C_{OUT} are the numbers of vehicles captured by the in-set and out-set point detectors, respectfully, at the end of the time interval.

Figure 5-2 shows an example of FATV at the I-210 and I-605 interchange. The blue arrows show the entering traffic, while the orange ones show the exiting traffic volume. The plots show the total number of egress and ingress vehicles captured by the VDS sets. The same VDS 773205 is part of the in-set and out-set for two different FATVs. In both figures, the inflow line closely follows the outflow line, which means the VDSs are working properly with limited possibilities of miscounting. The reason that the two lines are not perfectly aligned is because vehicles accumulate inside the volume, especially during congested periods.



(a) FATV with VDS 773205 in the in-set



(a) FATV with VDS 773205 in the out-set

Figure 5-2 FATV example demonstration

The FATV concept not only provides an automated validation method, but effectively groups sensors into an organized hierarchy. This hierarchy can then be used to map FATV groups to a spatial network for fusion with third-party data.

2.2. PROPOSED STRATEGY FOR VHD IN PEMS

Third-party data, which provides speed and travel time values along segments, naturally lends itself to delay estimation. The hybrid data method proved effective, yielding accuracy gains when fixed sensor groups (i.e., FATV) were selectively removed. Third-party data showed promising results for roadways with limited instrumentation. However, there are caveats. To summarize, the key conclusions are:

- **Third-party data can compensate for the loss of point-based sensors**
    - Possible to remove less critical sensors where flow changes are small
    - Not advisable to remove critical sensors at locations used for control, or where flow changes are large (e.g., fwy-fwy connector and interchanges)
    - More investigation is needed to identify key factors for sensor removal
- **Third-party data can *roughly* estimate delay on roadways with limited instrumentation**
    - VHD error was largely dependent on the accuracy of hourly flow profiles
    - Generic profile incurred severe error (>40%)
    - Measured profiles from nearby sensors incurred high, but less severe error (≈20%)

The strategic implications of these results are that hybrid data can increase the distance between point-based sensors while limiting the degradation of VHD results. This has the greatest implications for well-instrumented roadway segments with many point-sensors. However, areas with complex flow patterns or needs for traffic control still require point-based sensors, such as critical interchanges, added/removed lanes, and HOV/HOT lanes. In these cases, the probe-based third-party data may not provide enough spatial precision necessary for analysis and management.

While hybrid data can compensate for VDS loss in VHD estimation, replacing point-based sensor data entirely on roadways with limited instrumentation yielded relatively weak results. Estimating VHD using only third-party data on roadways with limited instrumentation using generic and measured flow profiles yielded a mean absolute error of 41.3% and 19.4%, respectively. This is substantially higher than the traditional and hybrid methods with point-based sensors at 12.7% and 3.4%, respectively. In general, VHD estimation on roadways with limited instrumentation is not recommended for any sensitive applications where precision is required. However, there may be applications where a rough value is useful, even with accuracy limitations. For example, a rough estimate of the overall delay could be used as justification for a more detailed investigation. In these cases, it may prove useful to include VHD from roadways with limited instrumentation in PeMS. In all cases, the provenance of the data should be maintained along with some expectation of its precision.

2.3. RECOMMENDATIONS

There are four main methods that were compared for estimating VHD:

- Traditional data and calculation
  - Uses point-sensor data only
  - Calculates delay over long freeway segments
- 3rd party + traditional calculation
  - Only possible when spatial reference systems match
  - Uses point-sensor data for flows and third-party data for travel times
  - Calculates delay over one connector, or a long freeway segment
- Hybrid calculation
  - Required when spatial reference systems do not match
  - Uses point-sensor data for flows and third-party data for travel times
  - Divides long freeway segments into cells for greater accuracy
  - Applies traffic theory to accommodate the distance between point-sensors
- Adjustments for limited instrumentation
  - Uses rough estimates for flows and third-party data for travel times

Table 5-1: Recommended Delay Calculation Method for Each Facility

	ML	HOV	Connectors	Ramps	Arterials
Traditional data and calculation		3rd party data not widely available			
3rd party + traditional calculation			Obtained good performance		
Hybrid calculation	Obtained best performance	Potential for the future			
No recommendation				Needs further work	Needs further work

Based on the analysis in this report, the recommended VHD estimation method depends on the infrastructure type and the data available. Recommendations are summarized in Table 5-1. For freeway mainlines, the best performance was achieved with the hybrid calculation. For HOV lanes, traditional traffic sensing methods must be used until third-party data become available that offer precision to reliably distinguish HOV lanes from mainline lanes. For connectors, good performance was obtained using third-party data combined with the traditional calculation. For ramps and arterials, further work is required.

3. THIRD PARTY DATA ROADMAP STRATEGY

A roadmap towards the implementation of a hybrid data collection strategy for Caltrans should advance the following broad goals:

- Reduce costs and increase coverage of traffic monitoring
- Provide a sound methodology for VHD estimation
- Enable a smarter deployment of point-based sensors

In addition, a path forward should:

- Leverage new and emerging technologies such as cloud computing
- Benefit from Caltrans' investment in data hub development in Connected Corridors
- Plan for integration with multiple data types and multiple providers

Cloud computing provides the opportunity to contract out the business of maintaining physical hardware and servers. It also allows for increased flexibility, scalability, and the ability to incorporate new technologies quickly.

A hybrid approach to data requires the ability to tap into multiple sources while also keeping track of quality control, and data provenance. Key technical challenges include:

- Networking and communications
- Data ingestion and quality control
- Spatial reference translation
- Data fusion and performance metric estimation
- Reporting and visualization

In addition to the technical challenges, any hybrid data solution will require ongoing support and maintenance. A good solution will support gradual experimentation and the ability to be shaped incrementally from direct experience with third-party data.

We recommend the following two-step roadmap:

<u>Step 1: Limited pilot</u>

A first step on the path towards implementation would consist of launching a limited pilot over one or several freeways that have excellent data and have been well studied. This choice will minimize the effort and risk associated with spatial reference translation.

If possible, it is beneficial to leverage any prior investments in experimentation platforms for multiple systems integration and processes that leverage third-party data, or data other than that procured through systems solely operated by Caltrans. The pilot system would operate in parallel and complement ATMS and PeMS.

The importance of continual maintenance of transportation asset data cannot be understated. Assets include the physical road network as well as the sensors providing the data. Each provider may have its own method for updating its representation of the road network and the delivery of its travel time estimates. Any hybrid system will need to solve the conflation problem—that is to project data from any source onto the desired domain of analysis to calculate and to report the performance metrics of interest. If possible, it is helpful to select a corridor in which the initial work to organize VDS into FATVs with detailed geospatial information has already been completed.

In this step, modifications to detector deployment strategy and the potential trade-offs of accuracy and cost could be evaluated in detail before being implemented on a wider scale. An additional goal would be to investigate details of performance metrics calculation involving such facilities as HOV lanes and ramps. Results will inform the potential expansion of hybrid data techniques.

This research project focused on the performance measure of delay. However, there are other needs for data, such as situational awareness and for real-time traffic management. A limited pilot would enable experimentation on a small scale to incorporate other metrics and the ability to visualize and display traffic information for other purposes.

Key research related tasks that could inform such a limited pilot are as follows:

- Create an initial set of freeways with high quality and reliable data.
- Pre-select sites in which the existing placement of loops or other point-detectors are likely to provide synergy with third-party data.
- Perform an initial FATV assessment of selected freeways to help determine more precise location information for sensors at freeway-freeway connectors.
- Finally perform a redundancy analysis to prioritize existing sensors relative to their marginal information content.

Step 2: Full-scale pilot in selected district

The second step we propose is a full-scale pilot in a selected district before Caltrans should elevate a hybrid traffic data collection strategy to a statewide practice.

Under this scenario, a fully operational traffic management system fed by a hybrid set of traffic data would be rolled out in the district's TMC. That traffic management system may either be an evolution of current ATMS software, a module thereof, or a brand-new implementation.

Innovations and lessons learned from Step 1 could be applied to provide new sets of metrics and new abilities to visualize and display traffic information. In addition to visualizing data from individual sensors, this system would have the capability to display traffic state and performance metrics based on a hybrid fusion of data from all available sources. Situational awareness could be based on a state estimator that takes all the data into account. This is a marked and important change, as it will be important to have the ability to switch among data vendors as legacy products evolve and new products are brought into the data marketplace.

The new traffic management system could initially be rolled out in parallel to the existing ATMS to ensure a smooth transition and the ability to roll back if needed.

Such an effort would demonstrate Caltrans' ability to evolve with changing technologies, and to leverage the benefits of cloud computing. In an increasingly interconnected world, the future of transportation management will require better and more complete data that can only be obtained through greater connectivity to the data feeds of private vendors as well as increased cooperation and collaboration with local stakeholders. A first step in this direction is to adopt computing tools and infrastructure that have already been tried and implemented at scale in the private sector.

In the past, a perceived risk of using hybrid data for traffic management systems was its dependence on external vendors. However, the new risk is that mobile devices are so prolific that drivers are now being influenced by the apps, and traffic management systems lack direct access to this influential, and useful, information.

Of course, data quality would need to be monitored on a continuous basis, and attention should be paid to costs when selecting a portfolio of data sources. This perspective applies also to the role and usage of traditional detectors as well. The deployment of a traffic management system that uses hybrid traffic data will provide the opportunity to start altering the selected district’s detector strategy. It would enable defining critical detectors that need continued maintenance and reassess future needs.

Experiences gained from this step on the implementation path should provide perspective on how to manage new risks and to benefit from new opportunities. Cutting-edge solutions adopted in the selected district can become a model of organizational excellence and be copied across California and around the world.

Bibliography

Ambühl, L. and Menendez, M. (2016) 'Data fusion algorithm for macroscopic fundamental diagram estimation', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2016.07.013.

Bayen, A. M., Sharafsaleh, P. E. and Patire, A. D. (2013) *Hybrid Traffic Data Collection Roadmap: Objectives and Methods*.

Caltrans (2012) *Statewide Mobility Performance Report*. Available at: http: //www.dot.ca.gov/trafficops/mpr/source.html (Accessed: 19 June 2019).

Caltrans (2020a) *PeMS Data Source | Caltrans*. Available at: http://pems.dot.ca.gov/ (Accessed: 6 November 2019).

Caltrans (2020b) *Quarterly Reports | Caltrans*. Available at: https://dot.ca.gov/programs/traffic-operations/mpr/quarterly (Accessed: 7 June 2020).

Center for Advanced Automotive Technology (no date) *Connected and Automated Vehicles*. Available at: http://autocaat.org/Technologies/Automated_and_Connected_Vehicles/ (Accessed: 19 June 2019).

Chen, X. Z. C. V. D. G. E. and Mei (2019) *Practices on Acquiring Proprietary Data for Transportation Applications*, *Practices on Acquiring Proprietary Data for Transportation Applications*. doi: 10.17226/25519.

CITILABS (2019) *CITILABS*. Available at: https://www.citilabs.com/software/streetlytics (Accessed: 25 October 2019).

Connected Corridors (2020) *I-210 Pilot | Connected Corridors Program*. Available at: https://connected-corridors.berkeley.edu/i-210-pilot-landing-page (Accessed: 17 June 2020).

HERE (2019) *HERE Technologies*. Available at: https://www.here.com/about-us (Accessed: 19 October 2019).

INRIX (2019) *INRIX*. Available at: http://inrix.com/about (Accessed: 21 October 2019).

Khaleghi, B. *et al.* (2013) 'Multisensor data fusion: A review of the state-of-the-art', *Information Fusion*. doi: 10.1016/j.inffus.2011.08.001.

Li, M., Chen, X. (Michael) and Ni, W. (2016) 'An extended generalized filter algorithm for urban expressway traffic time estimation based on heterogeneous data', *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. doi: 10.1080/15472450.2016.1153426.

Van Lint, J. W. C. and Hoogendoorn, S. P. (2010) 'A Robust and Efficient Method for Fusing Heterogeneous Data from Traffic Sensors on Freeways', *Computer-Aided Civil and Infrastructure Engineering*. doi: 10.1111/j.1467-8667.2009.00617.x.

Liu, J. *et al.* (2020) 'Urban big data fusion based on deep learning: An overview', *Information Fusion*, 53, pp. 123–133. doi: 10.1016/j.inffus.2019.06.016.

Ma, X. *et al.* (2017) 'Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction', *Sensors*. MDPI AG, 17(4), p. 818. doi: 10.3390/s17040818.

Ottaviano, F., Cui, F. and Chow, A. H. F. (2017) 'Modeling and data fusion of dynamic highway traffic', *Transportation Research Record*, 2644, pp. 92–99. doi: 10.3141/2644-11.

Patire, A. D. *et al.* (2015) 'How much GPS data do we need?', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2015.02.011.

Roess, R. P., Prassas, E. S. and McShane, W. R. (2011) *Traffic Engineering*.

Sabour, S., Frosst, N. and Hinton, G. E. (2017) 'Dynamic routing between capsules', *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), pp. 3857–3867.

Streetlight Data Inc. (2019) *Streetlight Data Inc.* Available at: https://www.streetlightdata.com (Accessed: 25 October 2019).

TomTom (2019) *TomTom*. Available at: https://www.tomtom.com/company/ (Accessed: 23 October 2019).

Treiber, M. and Helbing, D. (2002) 'Reconstructing the Spatio-Temporal Traffic Dynamics from Stationary Detector Data', *Cooperative Transportation Dynamics*.

Treiber, M., Kesting, A. and Wilson, R. E. (2011) 'Reconstructing the Traffic State by Fusion of Heterogeneous Data', *Computer-Aided Civil and Infrastructure Engineering*. doi: 10.1111/j.1467-8667.2010.00698.x.

Wang, Y. *et al.* (2019) 'Multi-Source Traffic Data Reconstruction Using Joint Low-Rank and Fundamental Diagram Constraints', *IEEE Intelligent Transportation Systems Magazine*. Institute of Electrical and Electronics Engineers, 11(3), pp. 221–234. doi: 10.1109/MITS.2019.2919529.

Wright, M. and Horowitz, R. (2016) 'Fusing loop and GPS probe measurements to estimate freeway density', *IEEE Transactions on Intelligent Transportation Systems*. doi: 10.1109/TITS.2016.2565438.

Wu, C. *et al.* (2015) 'Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization', *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 59, pp. 111–128. doi: 10.1016/j.trc.2015.05.004.

Zwet, van E. *et al.* (2003) 'A statistical method for estimating speed from single loop detectors'. Available at: https://drive.google.com/file/d/0B5wZ4dLpgONnT1NxbG9SQ21rR1k/edit.

Appendix A – VHD Calculation Error at Individual VDS Locations

For Task 4, the VHD calculation error caused by the removal of each FATV group is listed in the following table. The table has the errors calculated by both the traditional method and the hybrid method, and for all four evaluation scenarios. The absolute percentage error is calculated compared to the simulated ground truth values. The ‘VDS IDs of removed FATVs’ column has the IDs of the VDSs of the FATV group which is removed in the experiment, meaning the count and occupancy data of the VDSs are not available. The cross streets are the names associated with the VDS.

Table: Absolute percentage error of VHD calculation

VDS IDs of removed FATVs	Cross Street of the removed VDSs	Abs % Error using traditional method				Abs % Error using hybrid method			
		Before morning peak	Morning peak	Noon time	Afternoon peak	Before morning peak	Morning peak	Noon time	Afternoon peak
['717685', '772902']	CITRUS and PASADENA AVE	19.08	14.84	3.98	13.58	2.50	1.13	5.60	5.16
['772902', '717682']	PASADENA AVE and AZUSA 2	14.41	12.23	4.20	13.47	1.99	1.17	5.51	5.67
['717682', '717678']	AZUSA 2 and AZUSA 1	14.77	12.37	4.42	13.23	2.99	1.28	4.53	5.54
['717678', '717676']	AZUSA 1 and VERNON	15.54	13.16	2.85	13.46	2.42	1.97	5.50	5.93
['717676', '772888']	VERNON and ZACHARY PADILLA	13.78	12.70	2.87	13.69	2.66	1.86	5.10	5.52
['772888', '717675']	ZACHARY PADILLA and IRWINDALE 2	12.97	12.39	2.56	13.79	2.89	0.53	4.78	3.98
['717675', '717674']	IRWINDALE 2 and IRWINDALE 1	11.05	11.43	1.88	13.81	2.35	0.90	4.72	5.77
['717674', '772873']	IRWINDALE 1 and W/O IRWINDALE	9.19	10.51	2.69	13.53	3.45	2.14	4.96	6.50
['772873', '772858']	W/O IRWINDALE and SAN GABRIEL RIVER	9.09	11.41	2.24	13.28	2.76	2.04	4.86	6.50
['772858', '717673']	SAN GABRIEL RIVER and MOUNT OLIVE DR / 605	8.30	13.24	9.03	13.14	3.27	0.57	5.13	4.93
['717673', '769722']	MOUNT OLIVE DR / 605 and NB 605 TO WB 210 CON	7.53	12.22	7.15	13.14	3.91	0.50	3.66	4.24
['769722', '769702']	NB 605 TO WB 210 CON and HIGHLAND	7.53	12.00	4.77	13.14	4.66	0.65	2.87	4.30
['769702', '761374']	HIGHLAND and BUENA VISTA	9.82	17.01	13.26	14.50	3.77	0.09	4.14	5.00

['761374', '718210']	BUENA VISTA and MOUNTAIN AV	14.63	19.59	13.77	16.63	2.61	0.66	6.04	6.42
['718210', '761356']	MOUNTAIN AV and MYRTLE AV	14.15	20.71	13.85	22.36	4.13	0.42	4.29	5.31
['761356', '761342']	MYRTLE AV and HUNTINGTON 1	17.75	20.31	13.49	26.27	4.61	1.27	4.87	5.55
['761342', '773194']	HUNTINGTON 1 and E OF SECOND	14.00	16.68	7.76	23.12	4.46	0.82	3.26	4.09
['773194', '764146']	E OF SECOND and SANTA ANITA 2	11.12	14.70	4.27	20.66	3.77	1.19	2.80	3.67
['764146', '717669']	SANTA ANITA 2 and SANTA ANITA 1	12.19	15.30	6.00	21.46	4.55	0.30	2.33	3.79
['717669', '717664']	SANTA ANITA 1 and BALDWIN 2	14.19	16.81	7.99	23.31	3.93	0.23	2.25	4.41
['717664', '717663']	BALDWIN 2 and BALDWIN 1	10.87	14.72	3.96	20.29	3.72	1.12	4.00	4.64
['717663', '773179']	BALDWIN 1 and VAQUERO	7.36	12.43	0.89	16.78	4.02	1.01	4.17	3.79
['773179', '717661']	VAQUERO and MICHILLINDA	8.28	13.13	0.82	17.78	5.03	0.60	3.20	3.57
['717661', '717657']	MICHILLINDA and ROSEMEAD 2	6.46	12.22	1.39	16.90	5.33	0.83	2.78	3.63
['717657', '717653']	ROSEMEAD 2 and ROSEMEAD 1	5.08	12.14	4.59	17.55	5.39	0.31	3.12	3.54
['717653', '717649']	ROSEMEAD 1 and SIERRA MADRE V1	8.58	15.79	5.11	24.12	5.85	0.30	3.06	3.49
['717649', '717644']	SIERRA MADRE V1 and SAN GABRIEL	11.42	17.48	5.04	28.32	5.81	0.90	3.02	3.25
['717644', '717642']	SAN GABRIEL and ALTADENA	12.36	17.48	5.13	28.69	5.05	0.34	3.94	3.58
['717642', '717637']	ALTADENA and HILL	18.61	19.15	5.07	35.29	4.51	1.65	2.73	3.58
['717637', '717634']	HILL and LAKE 1	15.04	15.15	3.47	29.71	6.23	0.58	2.96	3.51