Developing an eScience Transportation Platform for Freeway Performance Analysis

Sa Xiao

A thesis
submitted in partial fulfillment of the
requirements for the degree of
Master of Science

University of Washington

2013

Committee:
Yinhai Wang
Ed D. McCormack

Program Authorized to Offer Degree:
Department of Civil and Environmental Engineering

University of Washington

Abstract

Developing an eScience Transportation Platform for Freeway Performance Analysis

Sa Xiao

Chairs of the Supervisory Committee:

Professor Yinhai Wang

Department of Civil and Environmental Engineering

While the exponential growth of data brings us tremendous opportunities of research, it also creates key challenges that will need to be tackled. As of 2012, we create approximately 2.5 exabytes of information each day, which equals the total amount of data stored on magnetic tape in 2001. The way people create, store, maintain, access, share, and utilize the data leads to a brand new outlook called big data. It motivates and inspires scientists and researchers to develop new infrastructure for better exploiting and exploring huge amount of multidisciplinary data. In 1999, Jon Taylor, the Director General of Research Councils in the UK, first introduced the term "eScience", which defines the novel generation of infrastructure that enables researchers from multidisciplinary areas collaborate with each other to achieve better, faster, and diverse research capabilities.

Inspired by the concept of eScience, the on-line transportation platform Digital Roadway Interactive Visualization and Evaluation Network (DRIVENet) is developed aimed at transportation data sharing, integration, visualization, and analysis. The major research goals for the DRIVENet system can be summarized in threefold. First, it provides the repository service to facilitate data sharing and integration. Second, one of the primary purposes DRIVENet serve is to

visualize the large sets of transportation data, helping users to perceive and understand the data. Third, the interactive and computational functions built in the DRIVENet system allow users to perform a variety of statistical analysis on multiple data sources, assisting users to draw meaningful inferences and make informed decisions. This research thus attempts to propose an innovative system architecture to address the aforementioned challenges, and develop an eScience approach to effectively utilize the existing data resources for transportation applications. Specially, a new approach that automates real-time freeway performance measurement is developed and implemented on DRIVENet, which further demonstrates the capability of DRIVENet in solving transportation problems. The new approach also provides quantitative evaluation of network-wide freeway performance to facilitate decision making in transportation operations and management.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I never would have completed this thesis without the love, support, and help from my supervisor, friends, and family. Now towards the end of my master's journey I would like to give my sincere acknowledgement to those who have helped me and influenced my work and life.

First of all, a big thanks to my supervisor Dr. Yinhai Wang for his patient supervision and profound guidance throughout my research. He has opened new doors in transportation engineering for me and constantly pushed and encouraged me to challenge myself. I am also truly grateful to Dr. Edward D. McCormack for serving on my thesis committee and to Ms. Kumiko Izawa for her valuable advice. Their extensive assistance and suggestions mean a lot to me. In addition, I greatly appreciate the financial support from Washington State Department of Transportation for the DRIVENet research.

I would like to express my deepest appreciation towards fellow students in STARLab. Particularly I am indebted to my senior colleagues Dr. Xiaoyue Liu and Dr. Runze Yu, who are not only my friends but important mentors. During these two years of study, I have learned so many things from you both academically and spiritually. Your integrity, honesty, and wisdom has been and will be inspiring and motivating me for a long time. I wish to thank Dr. Yegor Malinovskiy for his help in configuring R/Rserve. Additionally I extend my appreciation to Dr. Xiaoyue Liu for her great deal of time and efforts on improving my writing.

My hearty gratitude to my friends is beyond words. They make my life a miraculous journey, full of laugher, joy, and bliss. Special thanks goes to Ms. Jingren Gu (thank you for everything), Mr. Xudong Li, Mr. Chang Dou, and Mr. Felix Ye. I am grateful from the bottom of my heart for the way you support me, the care you give me, and all the good and bad moments we

have shared together. I also wish to thank my dear friends Ms. Xiaoyi Liu and Ms. Huizhong Guo for their understanding and unconditional support. My apologies for not being able to join your journey to San Francisco. A particular acknowledgement goes to Ms. Yitong Zhang and Ms. Xi Zhan, who took care of me and involved me in all kinds of activities in my first year at Seattle. I am very much thankful for invaluable friendship with Ms. Eve Zhao, Mr. Zhongqi Lu, Mr. Yaoyu Yang, Mr. Jingda Wu, Ms. Yi Pan, Ms. Xian Gong, my dear friends from Hong Kong University of Science & Technology, and those I could not mention here personally. My life would not have been so pleasant and meaningful without all my amazing friends.

Finally, and most importantly, I would like to thank my parents Ms. Huiqi Yan and Mr. Dong Xiao, who are always there for me. You have given me the greatest gifts of all, love, happiness, courage, and freedom of choice. I dedicate this thesis to you and love you with all my heart.

DEDICATION

To my family and all my dear friends.

Thank you.

Chapter 1: Introduction

1.1 General Background

While the exponential growth of data brings us tremendous opportunities of research, it also creates key challenges that will need to be tackled. As of 2012, we created approximately 2.5 exabytes (quintillion bytes) of information each day (Eaton *et al.*, 2012), which equals the total amount of data stored on magnetic tape in 2001 (Marcella *et al.*, 2002). The data can be generated from everywhere in any format, such as messages, social networking updates, videos, GPS locations, transaction records, etc. The study conducted by IDC Digital Universe demonstrates that world's information is more than doubling every 2 years, which is faster than Moore's Law (Gantz *et al.*, 2011). The way people create, store, maintain, access, share, and utilize the data leads to a brand new outlook called big data.

Big data motivates and inspires scientists and researchers to develop new infrastructure for better exploiting and exploring petabytes of multidisciplinary data. In 1999, John Taylor, the Director General of Research Councils in the UK Office of Science and Technology, first introduced the term "eScience" (Hey *et al.*, 2002). The concept of eScience captures the novel generation of infrastructure that enables researchers from multidisciplinary areas collaborate with each other to achieve better, faster, and more diverse research capabilities. In 2001, the United Kingdom government funded a £250 million 5-year eScience research project aiming to develop technologies, tools, and infrastructure to facilitate interdisciplinary collaboration, while U.S. government targets more than $200 million for big data projects in 2012 (Gianchandani, 2012). The CERN laboratory in Geneva, as a representative eScience research institute, conducts data-

oriented and computationally intensive experiments with the collaboration of more than 8000 researchers from over 100 worldwide institutions. A global eScience infrastructure, the LHC Computing Grid, has thus been built to distribute and analyze the huge amount of experimental data over the world to eventually realize the computational/data resource sharing (Hey *et al.*, 2005).

1.2 Research Goals

Inspired by the concept of eScience, the on-line transportation platform Digital Roadway Interactive Visualization and Evaluation Network (DRIVENet) was developed in 2008 (Ma *et al.*, 2011) aimed at data sharing, integration, visualization, and analysis. The system provides users with the capability to store, access, and manipulate data from anywhere as long as they have Internet connections. The major research goals for the DRIVENet system can be summarized in threefold. First, it provides the repository service to facilitate data sharing and integration. The existing data sources integrated amongst various organizations include roadway geometric data, loop detector data, Bluetooth data, INRIX speed data, incident data, weather data, freeway travel time, etc. Second, one of the major purposes DRIVENet serve is to visualize the large sets of transportation data, helping users to perceive and understand the data. Third, the interactive and computational functions built in the DRIVENet System allow users to perform a variety of statistical analysis on multiple data sources, assisting users to draw meaningful inferences and make informed decisions. The system benefits not only transportation practitioners and researchers but also the public by providing both historical and real-time transportation information and numerous performance measures in the broader context of an interdisciplinary framework.

1.3 Problem Statement

Despite many years of development, several challenging problems remained unsolved in the previous version DRIVENet 2.0. One critical issue is that the earlier versions have little geo-processing power, which makes it difficult to store, analyze, and manipulate geographic data. Previous solutions include manually recording series of spatial locations (latitude and longitude) for lines and polygon in relational database. However, this ad hoc method is inefficient, unreliable and not able to meet the needs of modeling complex spatial relationships.

Additionally, DRIVENet 2.0 has severe bugs and is vulnerable to intensive visits due to the incompatibility issues amongst the development tools. Google Web Toolkit (GWT) is one of the major tools adopted in this earlier version, which allows developers to write in Java and the GWT compiler translates Java code into JavaScript. Although GWT is a widely used tool to develop JavaScript front-end applications, it requires a steep learning curve and needs developers to keep up with new technologies very often. Huge amount of time and efforts are demanded for maintaining and updating the system because of rapidly changing features of GWT. Therefore, more productive and straightforward development process is desired to ensure the stability of such online platforms. Another concern induced by the Google Maps in DRIVENet 2.0 is the licensing model revision announced by Google, Inc. in early 2012 (Google, 2012). Only the first 2,500 geocoding web services will be offered free daily. Access to Google Maps will not be granted if a system continuously exceeds usage limits. Potential maintenance costs thus urge the transfer of DRIVENet system to a more flexible yet reliable alternative web-mapping products, such as OpenLayers and OpenStreetMap (OpenLayers, 2013; OpenStreetMap, 2013). These led to the development of DRIVENet 3.0 to be described in this thesis.

1.4 Scope of Study

Addressing the aforementioned challenges is especially critical during the new framework design of DRIVENet system. Meanwhile, although DRIVENet conceptually provides an eScience platform for data-driven discoveries and decision making, it is still unclear how well it performs in reality. In this research, automating real-time freeway performance measurement is selected as a case study to test such functionality. Considering the fact that freeway performance analysis involve complicated interactions among geometric, environmental, political, behavioral, and technological features, it is an ideal example to examine and demonstrate the capability of DRIVENet from an eScience perspective.

The remainder of the thesis is organized as follows: Chapter 2 gives an overview of the state-of-the-art in transportation web-based GIS systems, followed by background on freeway performance measurement in Chapter 3. Chapter 4 describes the research originality, more specifically, the innovative system architecture adopted by DRIVENet, followed by an elaboration on the data warehouse and system implementation. In Chapters 5 and 6, the author presents an application of DRIVENet, automating real-time freeway performance measurement, with discussion of proposed methodologies and implementation results. Finally, Chapter 7 concludes the thesis and offers future research directions.

Chapter 2: State of the Art

During the past few years, much research in Intelligent Transportation System (ITS) has been focusing on developing web-based transportation information system, aiming to be user-friendly and real-time. In this section, several representative online transportation systems related to DRIVENet are presented.

2.1 Freeway Performance Measurement System (PeMS)

Established in 1998, PeMS is a freeway performance measurement system jointly developed by the University of California, Berkeley, California Department of Transportation (Caltrans), and the Partners for Advanced Transportation Technology (PATH). With the support from Caltrans and local agencies, the system provides various traffic data sources including traffic detectors, census traffic counts, incident logs, vehicle classification data, toll tag based data, roadway inventory, etc. These traffic data have been automatically collected and archived for over ten years and real-time information are updated from over 25,000 detectors (Chen *et al.*, 2001; Chen *et al.*, 2003).

As the critical component of Caltrans performance measurement system, PeMS shown in *Figure 2-1* provides a variety of freeway evaluations in terms of speed, occupancy, travel time, vehicle miles traveled, vehicle hours traveled, vehicle hours of delay, etc. The success of PeMS on freeway triggers the development of arterial performance evaluation. Following the basic principle of PeMS, Arterial Performance Measurement System (APeMS) was then implemented to estimate intersection travel time, control delay, and progression quality on arterials every 5 minute, using mid-block loop detectors (Tsekeris *et al.*, 2004; Petty *et al.*, 2005). Different from the open

availability of PeMS, APeMS has designated usage for stakeholders and is not accessible by the public.

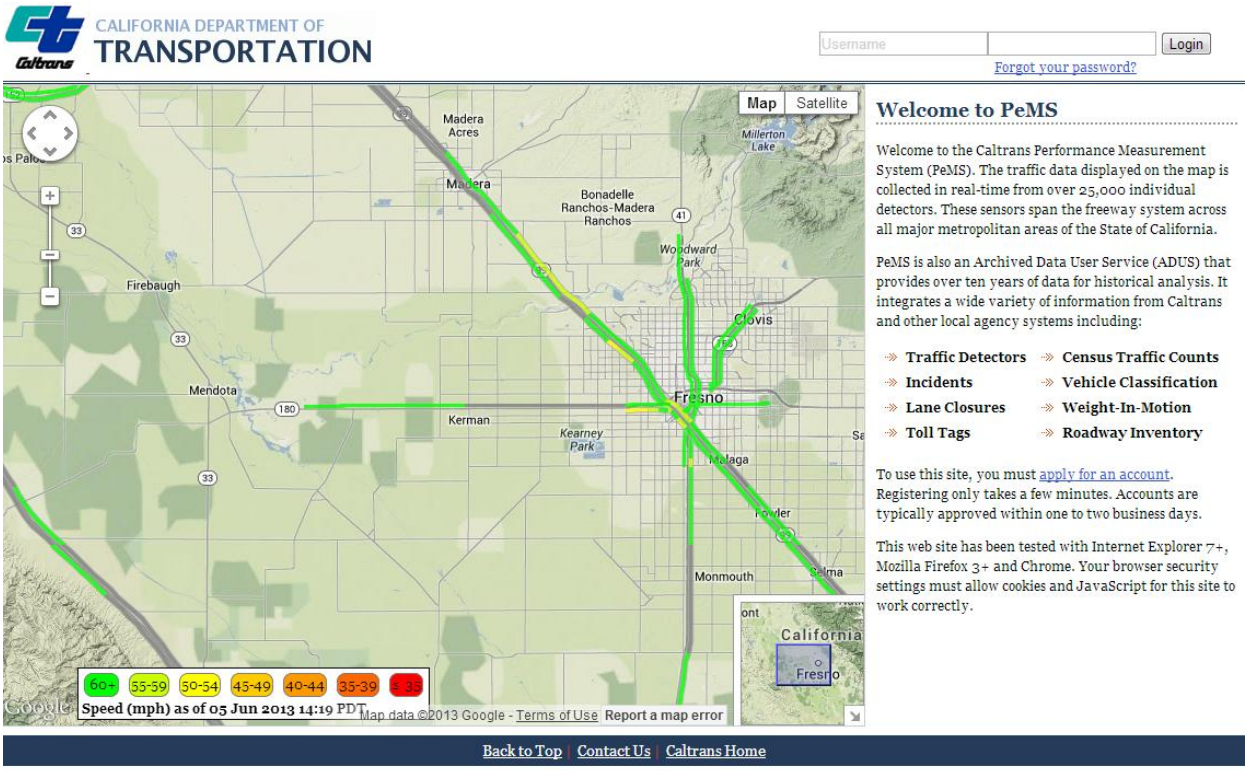


Figure 2-1 Screenshot of PeMS

2.2 Regional Integrated Transportation Information System (RITIS)

RITIS is an automated data archiving and integration system developed by the Center for Advanced Transportation Technology Laboratory (CATT Lab) at the University of Maryland. As one of most representative online transportation systems, RITIS targets to improve transportation safety, efficiency, and security by fusing and mining the transportation-related data in Maryland, Virginia, and the District of Columbia. The system provides both real-time and historical data to users with access credentials, including incident, weather, radio scanners, and other sensors.

Numerous visualization and analysis tools are developed to enable the interactive exploration and analysis of performance measures from data archival. DOT or public safety employees could possibly use the RITIS service by applying online. The system is not accessible to the general public (CATT Lab, 2012).

2.3 Portland Oregon Regional Transportation Archive Listing (PORTAL)

Originally established in 2004 with simple user interface and single data source – freeway loop detector data, PORTAL shown in *Figure 2-2* has improved significantly over the past eight years. In addition to the loop detector data from the Portland-Vancouver metropolitan region, now PORTAL 2.0 archives about one-terabyte transportation data including weather data, incident data, freight data and transit data. The system takes advantages of Adobe Flash and Google Maps technologies to display transportation data spatially. Additionally various graphical and tabulated performance information are available on the website, such as incident reports, transit speed map, traffic count, vehicle miles traveled, and vehicle hours traveled (Tufte *at el.*, 2010).

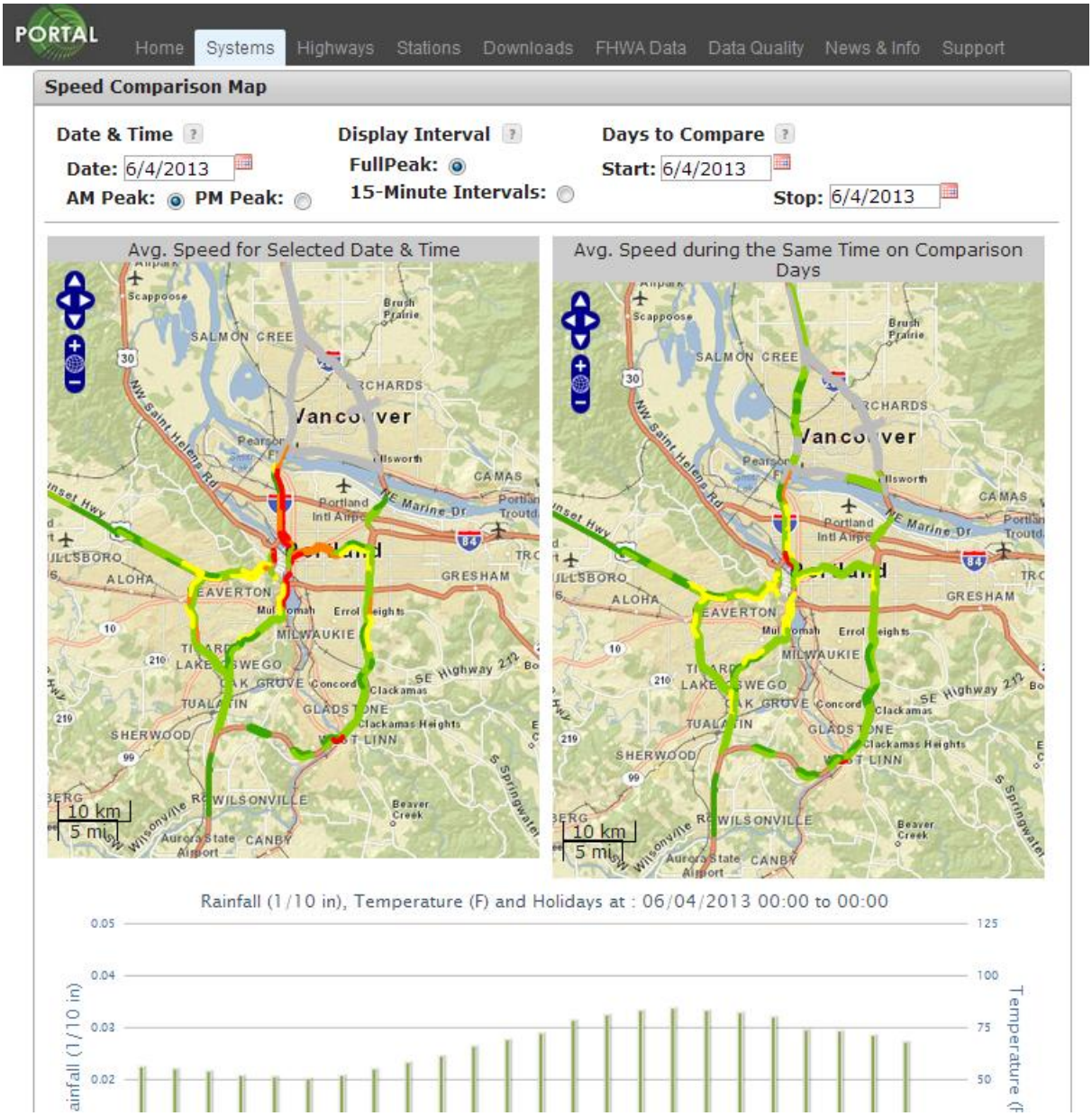


Figure 2-2 Screenshot of PORTAL 2.0

2.4 Freeway and Arterial System of Transportation Dashboard (FAST)

FAST dashboard, released online in September 2010 (http://bugatti.nvfast.org), is a web-based system developed to control and monitor the traffic in the Las Vegas and Nevada Metropolitan

areas (Xie *et al.*, 2012). In collaboration with the Nevada Department of Transportation, the system collects and archives real-time traffic data retrieved from loop detectors, radar detectors, and Bluetooth sensors deployed on freeways and ramps. Traffic data including lane occupancy, volume, and speed, are further processed as the major source for performance measurement. Also integrated in the system are incident data in the report format collected from the general public, and weather data shared by Nevada DOT Road Weather Information System.

The performance measures FAST dashboard uses includes average speed, traditional travel time performance measure, delay volume, and temporal and spatial extension of congestion. Meanwhile, the website is updated every 1-minute to display the real-time traffic map, as shown in *Figure 2-3*. By ensuring the delivery of timely and accurate information to traffic managers, operators, and planners as well as the general public, FAST dashboard significantly enhances the interchangeability of traffic data, helps improve the freeway and arterial system, and optimize the operation strategies in southern Nevada region.

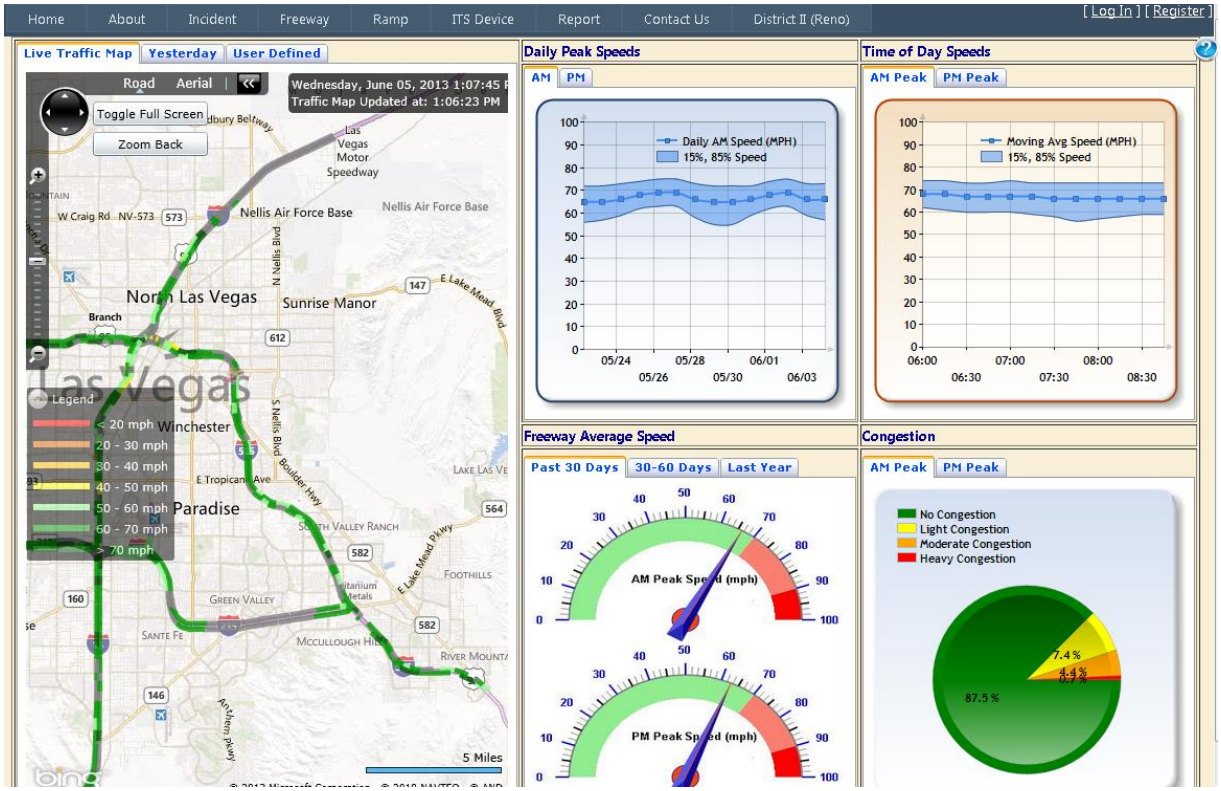


Figure 2-3 Screenshot of FAST Dashboard

Chapter 3: Background for Freeway Performance Analysis

This chapter gives readers the necessary background information for understanding freeway performance analysis in this thesis. It first introduces FREEVAL – 2010, a computational engine that implements Highway Capacity Manual (HCM) 2010 methodologies. The second part of the chapter describes the concept of Level of Service. In the end, it presents speed-density model as well as *K*-means Clustering used in this study.

3.1 FREEVAL – 2010

FREEVAL (FREeway EVALuation) is a computational engine built in the Microsoft Excel worksheet environment, to fully implement the computation for freeway facilities performance (Rouphail *et al.*, 2011). The computational procedure in FREEVAL implemented the methods presented in HCM 2010 Chapters 11, 12, and 13, for basic freeway segments, weaving segments, and merge and diverge segments, separately. The freeway facility with up to 70 analysis segments and 24 15-min time intervals are allowed to be queried at one time. In addition, some special requirements need to be satisfied due to the limitations of HCM methodologies. For instance, the temporal and spatial boundaries of analysis domain should have a demand-to-capacity ratio *d/c* less than 1. Before FREEVAL, most of the analysis were performed manually, which was time-consuming, inefficient, and sometimes unrealistic. FREEVAL enables the analytical automation as long as all necessary input requirements are fulfilled, which greatly reduces the cost and boost the productivity.

However, the success of automated computation provided by FREEVAL triggers other questions, e.g. how can we take advantage of huge volume of available transportation data to further avoid on-site data collection and manual data input? In Chapters 5 and 6, the proposed innovative data fusing techniques is aimed to solve this problem by presenting a prototype study on automating freeway performance measurement.

3.2 Level of Service

Level of service (LOS) is the most important and fundamental concept introduced by the Transportation Research Board (TRB) Highway Capacity Committee in early 1963. With the HCM evolvement over the past 50 years, LOS is now defined as a qualitative service measure of traffic operational conditions experienced by travelers with specific environmental characteristics. LOS simplifies the complex numerical results and further categorizes service condition into six levels from **A** to **F**, in which **A** represents the best condition while **F** describes traffic breakdown/failure.

Density is selected as the performance measure to define LOS on a basic freeway segment for three reasons: (1) Speed is insensitive to flow rates between 1,000 *pc/h/ln* to 1,800 *pc/h/ln* as shown in *Figure 3-4*; (2) Density naturally describes the headway between vehicles in traffic stream, which further reflects the ability to change lanes. (3) Density, more importantly, is sensitive to traffic flow rates. Therefore, LOS is primarily determined by traffic density for freeway facilities.

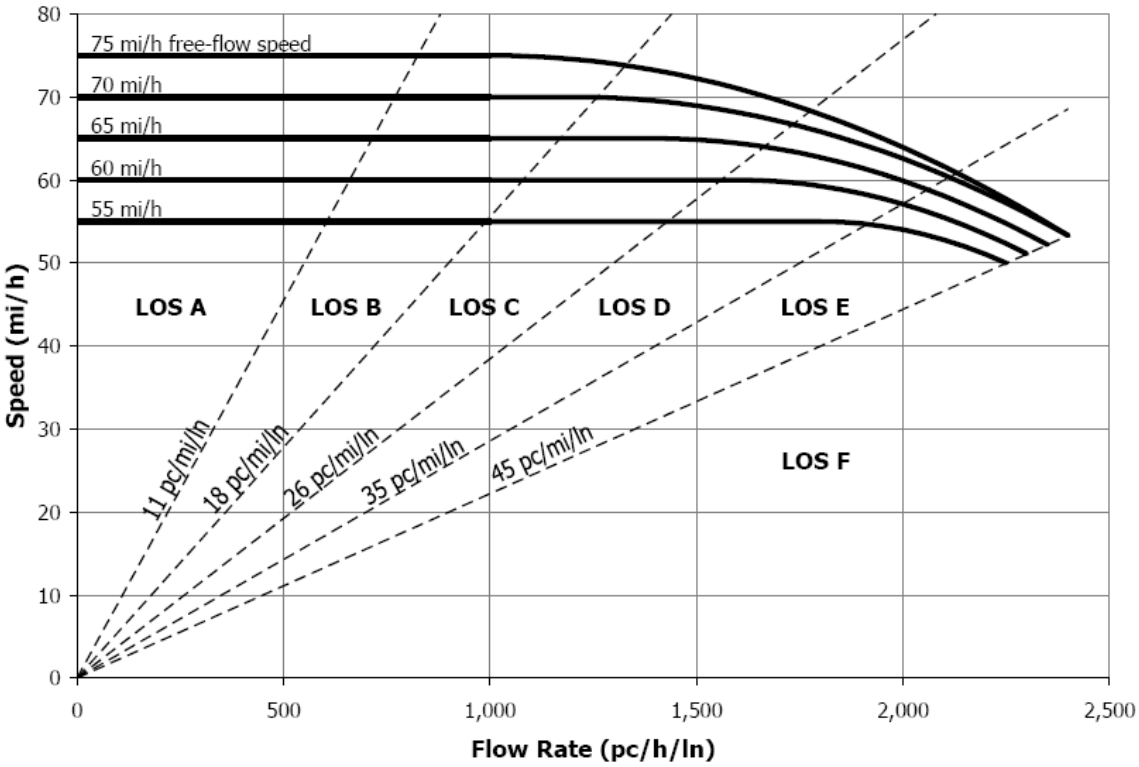


Figure 3-4 Speed-Flow Model for Basic Freeway Segments (HCM, 2010)

HCM defines the freeway LOS with the letters **A** through **F** as follows (HCM, 2010):

LOS **A**. LOS **A** describes roadway performance under the free flow conditions. Travelers have full freedom to choose speed and change lanes. The level of mobility and comfort travelers perceived is extremely high. It could reduce the effects of traffic breakdowns or incidents to the greatest extent.

LOS **B**. LOS **B** describes operations at or near the free-flow speeds. Travelers are slightly restricted to maneuver in the traffic stream compared to LOS A. The perceived level of mobility and comfort is still high.

LOS **C**. LOS **C** represents operations at or near the free-flow speeds. Travelers are noticeably restricted to change lanes and operate. More attention is required at the travelers' side. The level of mobility and comfort experienced by travelers remains at a reasonable level.

LOS **D**. LOS **D** represents the level at which speed drops with increasing flows. Travelers are more noticeably restricted to maneuver in the traffic stream. The impact of incidents or breakdown is severe, since there is little space to absorb disruptions.

LOS **E**. LOS **E** describes the operation at capacity. The state is vulnerable to any minor disruptions. The level of mobility and comfort travelers perceived is low.

LOS **F**. LOS **F** represents a traffic breakdown with low speed and little maneuverability. Facility undergoes considerable delay. In general, LOS **F** appears at facilities having more demand than capacity.

3.3 Speed-density model

The fundamental relationships of traffic flow describe the linkage amongst traffic characteristics of space mean speed v, density k, and flow q as follows:

Equation 3-1

$$q = k \times v$$

The quantitative relationships between any two of the three variables are of general interest. In 1935, Greenshields proposed the linear speed-density model illustrated in *Equation 3-2* and *Figure 3-5* (Greenshields, 1935).

Equation 3-2

$$v = v_f\left(1 - \frac{k}{k_j}\right)$$

Where v_f represents free flow speed and k_j represents the jam density.

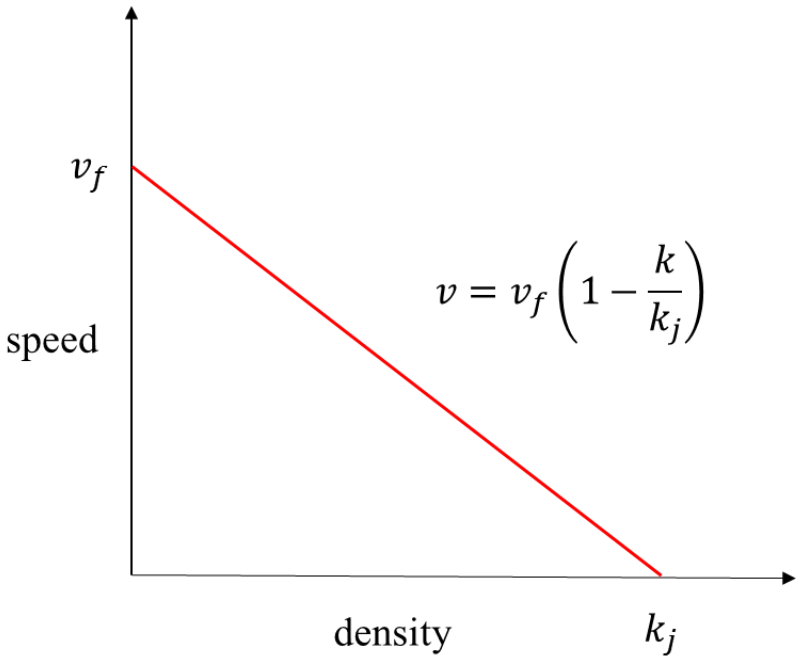


Figure 3-5 Greenshields Model

Although equations greatly simplify the problem elegantly, in reality the linear relationship hardly describes empirical observations well. Hence, a variety of single-regime models has been

proposed as listed in *Table 3-1*, in which v_f is free-flow speed, k_j represents jam density, v_o is optimal speed, and k_o is optimal density.

Single-regime Model	Function
Greenshields Model (1935)	$v = v_f(1 - k/k_j)$
Greenberg Model (1959)	$v = v_o \ln(k_j/k)$
Underwood Model (1961)	$v = v_f \, e^{-\frac{k}{k_o}}$
Northwestern Model (1967)	$v = v_f \, e^{-\frac{1}{2}(\frac{k}{k_o})^2}$
Drew Model (1968)	$v = v_f \left(1 - (\frac{k}{k_j})^{n+\frac{1}{2}}\right)$
Pipes Model (1967)	$v = v_f \left(1 - (\frac{k}{k_j})^{n}\right)$

Table 3-1 Single Regime Models

Recognizing the inability of single-regime models to fit the empirical observations, Edie first introduced two-regime model using Underwood's model for free-flow regime and Greenberg model for congested regime (Edie, 1961) as follows:

Equation 3-3

$$v = \begin{cases} 54.9 \exp(-k/163.9) \; for \; k \leq 50 \\ 26.8 \ln(162.5/k) \; for \; k > 50 \end{cases}$$

Following the idea of multi-regime models, Drake et al. (1965) proposed two two-regime models and one three-regime model based on the single-regime models listed in *Table 3-1*:

Equation 3-4

$$v = \begin{cases} 60.9 - 0.525k \; for \; k \leq 65 \\ 40 - 0.265k \; for \; k > 65 \end{cases}$$

Equation 3-5

$$v = \begin{cases} 48 \; for \; k \leq 35 \\ 32 \ln(145.5/k) for \; k > 35 \end{cases}$$

Equation 3-6

$$v = \begin{cases} 50 - 0.098k \; for \; k \leq 40 \\ 81.4 - 0.913k \; for \; 40 \leq k \leq 65 \\ 40 - 0.265k \; for \; k \geq 65 \end{cases}$$

However, the number of regimes and breakpoints in multi-regime models to be chosen is highly dependent on the engineering judgment, which is largely subjective and unscientific. To address this issue, Sun *et al.* (2005) proposed a methodology to automate the multi-regime regression from the data mining perspective. In general, traffic flow demonstrates two patterns: free flow and congested flow. Most of the time, there appears to be a third pattern called transition flow. In Sun *et al*. (2005)'s method, k-means clustering is adopted to naturally partition the empirical observations into clusters. Then, single-regime models are applied to fit each cluster and determine the breakpoints automatically.

3.4 *K*-means Clustering

K-means clustering (MacQueen, 1967; Steinhaus, 1957) is an unsupervised machine learning

approach that segments n observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ into k clusters $\mathrm{S} = (S_1, S_2, \ldots, S_k)$, which seeks to minimize the following objective function:

Equation 3-7

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2$$

In which $\boldsymbol{\mu}_i$ is the mean value of points in cluster S_i

The cluster analysis is a good method to produce relatively homogeneous groups of observations based on selected features. The most common algorithm to realize *K*-means clustering is the iterative refinement technique given as follows:

1. ***Initialize****: randomly select k points as cluster centers*
2. ***Repeat****:*
   a) *Assign each data point to the closest mean*
   b) *Update each cluster mean to be the average of its assigned points*
3. ***Stop****: when the assignments no long change*

Chapter 4: DRIVENet 3.0 Framework

This chapter is organized as follows: it first proposes the novel DRIVENet architecture design; then it presents the data warehouse and available data sets; in the end, implementation and user interface design are demonstrated.

4.1 DRIVENet System Architecture

The new system adopts the "thin-client and fat server" architecture with three basic tiers of web application, i.e. presentation tier, logic tier, and data tier, as showed in *Figure 4-6*. Presentation tier includes the user interface terminal via which users interact with the application. Logic tier, which is also called computational tier, is the core component of DRIVENet system. It performs computations in assisting customized analysis and decision making based on users' interactive input. Analytical tools developed include incident-induced delay forecasting using deterministic queuing theory (Yu *et al.*, 2011), Bluetooth-based pedestrian trajectory re-construction (Malinovskiy *et al.*, 2012), GPS-based truck performance measure (Ma *et al.*, 2011), etc. Data tier organizes and supports data requested for analysis. Normally the client handles the user interface while the server is responsible for the data. The significant difference between "thin-client and fat server" and "fat-client and thin server" is the shifted responsibility for the logic/computational Tier (Lewandowski, 1998). In fat server systems, the server fully takes over the logic/computation tier while the client only hosts the presentation tier for displaying user interface and dealing with user interaction.

There are three reasons to adopt a thin-client architecture: First, no plug-in and installation

is required at the client side except a basic browser, which ensures compatibility to the greatest extent. Considering the fact that the system is designed for customers with constrained network functions, minimal requirements at the client side is most desirable. Second, there is less security concern since all the data and computational tasks are manipulated and performed at the server side, while the client is only responsible for user interaction and results presentation. Third, mature frameworks for building thin client web application could be re-used to boost development productivity. For example, Vaadin is a Java framework that supports server-driven programming model (Vaadin, 2013). Since the coding is mainly based on Java, there is no need to learn other technologies such as JavaScript, potentially leads to less bugs and learning overhead. However, thin-client architecture does have its drawbacks. One major disadvantage is that the performance of system solely depends on the server and excessive user requests would greatly affect system efficiency. This can be remedied nowadays with the continuous advancement of cloud computing technologies such as Amazon Web Service, where the cloud servers are fully utilized to improve system performance.

The data communication flows in the DRIVENet system could be summarized as follows:

1. *The end-user sends an HTTP(S) request to the web server.*
2. *The web server looks into the request and retrieve the related data information from data warehouse.*
3. *The warehouse sends back the requested data and the web server performs the computational tasks using either the built-in analytical tools or external statistical modules provided by R Server.*
4. *If geospatial analysis is involved, the web server will connect to the OpenStreetMap Server and request the map.*
5. *Analysis results as well as the map are then returned to the client. Web browser displays the results or visualizes the returned objects on the map.*

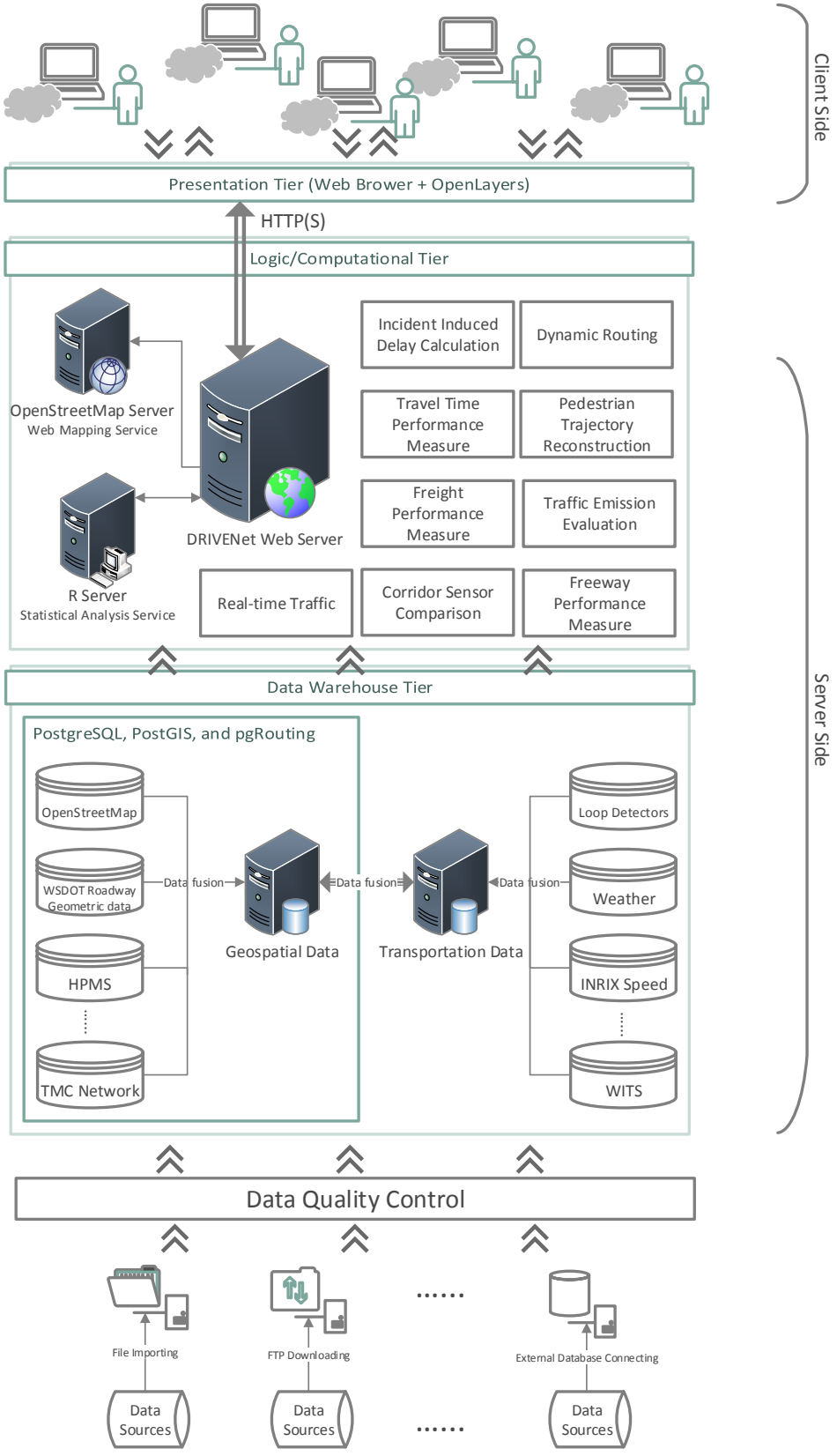


Figure 4-6 DRIVENet 3.0 Architecture

4.2 DRIVENet Data Warehouse

To achieve the integration and interoperability of various data sources, a DRIVENet data warehouse is designed and built for big data analysis. A variety of data are archived, replicated, transformed, and integrated in the data warehouse by web downloading (FTP, HTTP, or SOP), XML transferring, flat file exchanging, or direct collecting. These data sources are maintained by different agencies as indicated in *Table 4-2*. One reason that DRIVENet warehouse backs up data collected by cooperated agencies is that some of them may only keep the data for a limited time period due to storage constraints. Archived data provides historical information on the transportation system operations. Additionally, being an essential online tool, DRIVENet provides convenient and timely access to numerous data to support large scale analysis and decision making. The warehouse with integrated data further enables users to explore interdisciplinary relationships among transportation, environment, and human behaviors temporally and spatially. Users could either use the analytics modules in DRIVENet or directly download raw data for further analysis. Some example datasets are explained in the following subsections.

Data Source	Description	Coverage	Agency
Inductance Loop Data	Volume, occupancy, speed, and vehicle type	Washington State	WSDOT
Incident Data	Incident locations/types	Washington State	WSP
Surveillance Video	Roads and highways video	Puget Sound Area	WSDOT
Weather Data	Temperature/wind speed	Washington State	NOAA
Inductance Loop Data	Volume, occupancy, speed, and vehicle type	Bellevue	City of Bellevue
Sensys Data	Vehicle speed/volume	Seattle	City of Seattle
Inductance Loop Data	Second-by-second event data	Lynnwood	City of Lynnwood
Truck GPS Data	Freight movement data	Puget Sound Area	Commercial fleet companies
Speed Data	Based on vehicle GPS	Washington State	INRIX

Bluetooth Data	Travel time and speed	UW Campus, SR-520, I-90	STARLab, WSDOT
Geometric Factors	Shoulder width, number of lanes, lane width, etc.	Washington State	WSDOT
Freeway Data	Alerts, cameras, travel time	Washington State	WSDOT
Border Crossings	Wait time	I-5, SR-543, SR 539, SR-9	WSDOT
Mountain Pass Conditions	Weather, temperature, and conditions	Washington State	WSDOT

Table 4-2 Data Sources

4.2.1 Roadway Geometric Factors

WSDOT GIS and Roadway Data Office (GRDO) produces and maintains the GeoData Distribution Catalog online at http://www.wsdot.wa.gov/mapsdata/geodatacatalog/. The geospatial data in the format of ESRI Shapefile is available to the general public, promoting data exchange and data sharing. Various roadway geometric datasets are available, including number of lanes, roadway widths, ramp locations, shoulder widths, surface types, etc. State route ID and locations marked by mileposts and accumulated mileage are also included in the WSDOT linear referencing systems.

4.2.2 Loop Detectors

WSDOT deploys thousands of inductance loop detectors on freeway and highway networks in Washington State. Most of the loop detectors are set as single loop, providing real-time volume and occupancy aggregated every 20 seconds. Dual loop detectors comprises two paired single-loop detectors separated by several meters, used to measure speed and vehicle length. DRIVENet periodically archives and maintains both single and dual loop detector data from WSDOT. There are a total of 9729 single loops and 3671 dual loops included in the DRIVENet database. Being the main information source for traffic operations and decision making, loop data quality is a

critical issue as the loop malfunction and/or sensitivity level shift can result in significant detection bias. Data quality control strategies developed by STARLab are applied to ensure data quality (Wang *et al.*, 2009).

4.2.3 Washington Incident Tracking System (WITS)

Traffic incident data is collected and maintained by Washington State's Incident Response (IR) Team in the Washington Incident Tracking System (WITS). WITS includes majority of incidents happened on freeways and Washington State highways, which totaled 550,376 by March 2013. For each incident, Washington State IR team logs details such as incident location, notified time, clear time, closure lanes, etc. DRIVENet team obtained the WITS datasets from 2002 to 2013 and integrated them into the DRIVENet database. Using the methodology developed by Yu and Wang (Wang *et al.*, 2008; Yu *et al.*, 2011), incident-induced travel delays are further calculated and visualized in the DRIVENet system as showed in *Figure 4-7*.

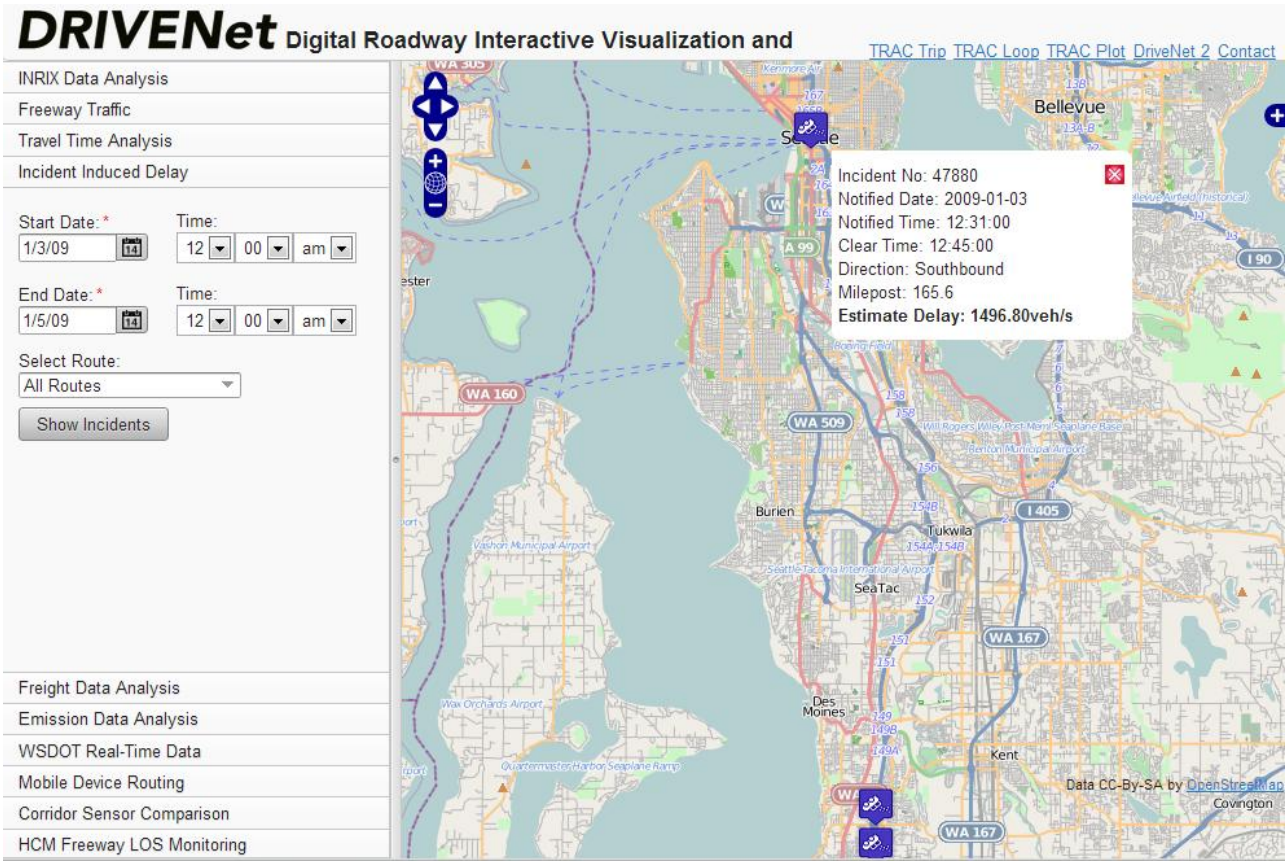


Figure 4-7 Incident Induced Delay

DRIVENet team recently retrieved INRIX speed datasets for 2008 through 2012 from WSDOT that purchased the license. As one of the major companies that produces traffic, INRIX analyzes and computes traffic information mainly based on measurement from GPS devices and loop detectors. The produced speed data was aggregated into 5-minute intervals for 2008, 2009, and 2010 and into 1-minute intervals for 2011 and 2012. It covers almost the entire State's roadway network, including freeway, highways, and most arterials and side streets. Considering that most of traffic evaluation in the previous DRIVENet system is based on loop detector data, the INRIX speed datasets are a great complement to the system. Traffic Message Channel (TMC), a common industry convention developed by leading map vendors, is adopted by INRIX as their base

roadway network. Each unique TMC code is used to identify a specific road segment. For example, in *Table 4-3*, TMC *114+0509* represents the WA-522 road segment with start location (*47.758321, -122.249705*) and end location (*47.753417, -122.277005*). However, since WSDOT follows a linear referencing system on the basis of mileposts, it poses challenges to match the two different roadway layouts for data fusion.

TMC	road	direction	intersection	county	zip	start point	end point	miles
114+05099	522	EASTBOUND	80th Ave	KING	98028	47.758321, -122.249705	47.755733, -122.23368	0.768734
114-05095	522	WESTBOUND	WA-523/145th St	KING	98155	47.753417, -122.277005	47.733752, -122.29253	1.608059

Table 4-3 INRIX Speed Data

4.2.4 Trucking GPS Data

DRIVENet team periodically and automatically fetches and imports GPS truck data collected by trucking companies into the data warehouse for freight performance measurement via an FTP connection. UW, WSDOT, and Washington Trucking Associations (WTA) signed contracts with three GPS vendors to acquire the data. *Table 4-4* provides general information on the data obtained from each vendor (Ma *et al.*, 2011). Common variables such as longitude, latitude, truck ID, travel heading, and timestamp are included in each datasets. Due to privacy concerns, vendor information is masked in this thesis.

Vendors	Average Daily Records	Total trucks per day	Frequency of Reads (min)	Data Type
Vendor A	94,000	2,500	5 ~ 15	In-car GPS with a cellular connection
Vendor B	12,000	25	0.5	In-car GPS with cellular connection

Vendor C	3,000	60	1 ~ 5	GPS mobile phone

Table 4-4 GPS Vendors (Ma et al., 2011)

4.3 System Implementation

As mentioned in the previous section, DRIVENet architecture has been re-designed to meet challenges. To reduce costs and boost productivity, multiple open source products are utilized for the system implementation. Relying on open source products, the DRIVENet team not only takes advantages of code-sharing and collaboration with a broad community of developers, but also contributes to open source projects. Core open source products combined into DRIVENet system are explained in the remainder of this section.

4.3.1 OpenStreetMap and OpenLayers

OpenStreetMap (OSM) is a collaborative project to create a comprehensive worldwide map that is free to use and editable (Haklay *et al.,* 2008). With the outlook that geospatial data should be freely accessible to the public, the OSM project was established by University College London in July 2004 and treated as one of the most prominent and famous examples of Volunteered Geographic Information, the concept introduced by Goodchild (2007, 2008). The process of maintaining OSM data is described as crowdsourcing which is also being used by other commercial companies such as Google and TomTom. The crowdsourcing, a term defined by Brabham as “online and distributed problem-solving and production model”, distributes the labor-intensive tasks to large groups of users and allows volunteers to create and update geospatial data on the Internet. By January 2013, OSM has over one million registered contributors and 20,000 active users worldwide and the number keeps rising dramatically (Wood, 2013). Additionally,

OSM obtained strong support from commercial companies as well as governments. For instance, Yahoo Maps made their vertical aerial imagery available to OSM as a backdrop for map production in 2006 and Microsoft Bing Maps donated part of its satellite imagery to the OSM in 2010 (Bing Blogs, 2010).

One major reason for DRIVENet to choose OSM is its low cost compared to commercial datasets as well as its data sharing nature. With the Open Data Commons Open Database License (ODbL), developers are free to use, distribute, and modify the OSM data as long as OSM and its contributors are credited (OpenStreetMap, 2013). On the one hand, using OSM to replace Google Maps helps DRIVENet avoid potential charges by Google, Inc in the future that might eventually prevent the project from growing. On the other hand, with the theme of eScience, DRIVENet prefers open source products over commercial ones, which could help share ideas, drive innovation, and boost productivity for the entire community.

High-resolution and qualitative geographic information as showed in *Figure 4-8* makes OSM an appealing replacement of Google Maps. Recent research confirms the good quality of OSM and its capability to compete against professional geodata, especially for urban areas. Zielstra and Hochmair (2011) used commercial datasets NAVTEQ and TeleAtlas as well as freely available dataset TIGER/Line to quantitate the coverage of OSM in the United States. The results indicate that "there is strong heterogeneity of OpenStreetMap data for the US in terms of its completeness". Similar study has been done by Zielstra and Zipf in 2010 for Germany (Zielstra *et al.*, 2010). The paper states that some projects already replace proprietary data with rich OSM data in larger cities. In U.K, Haklay (2010) performed a comparison using the Ordnance Survey (OS) Meridian dataset by evaluating accuracy, completeness, and consistency of it position and attributes. The analysis

reached the conclusion that "OSM information can be fairly accurate" with the positional accuracy of about 6 meter, and an approximately 80% overlap of motorway objects compared to OS datasets.

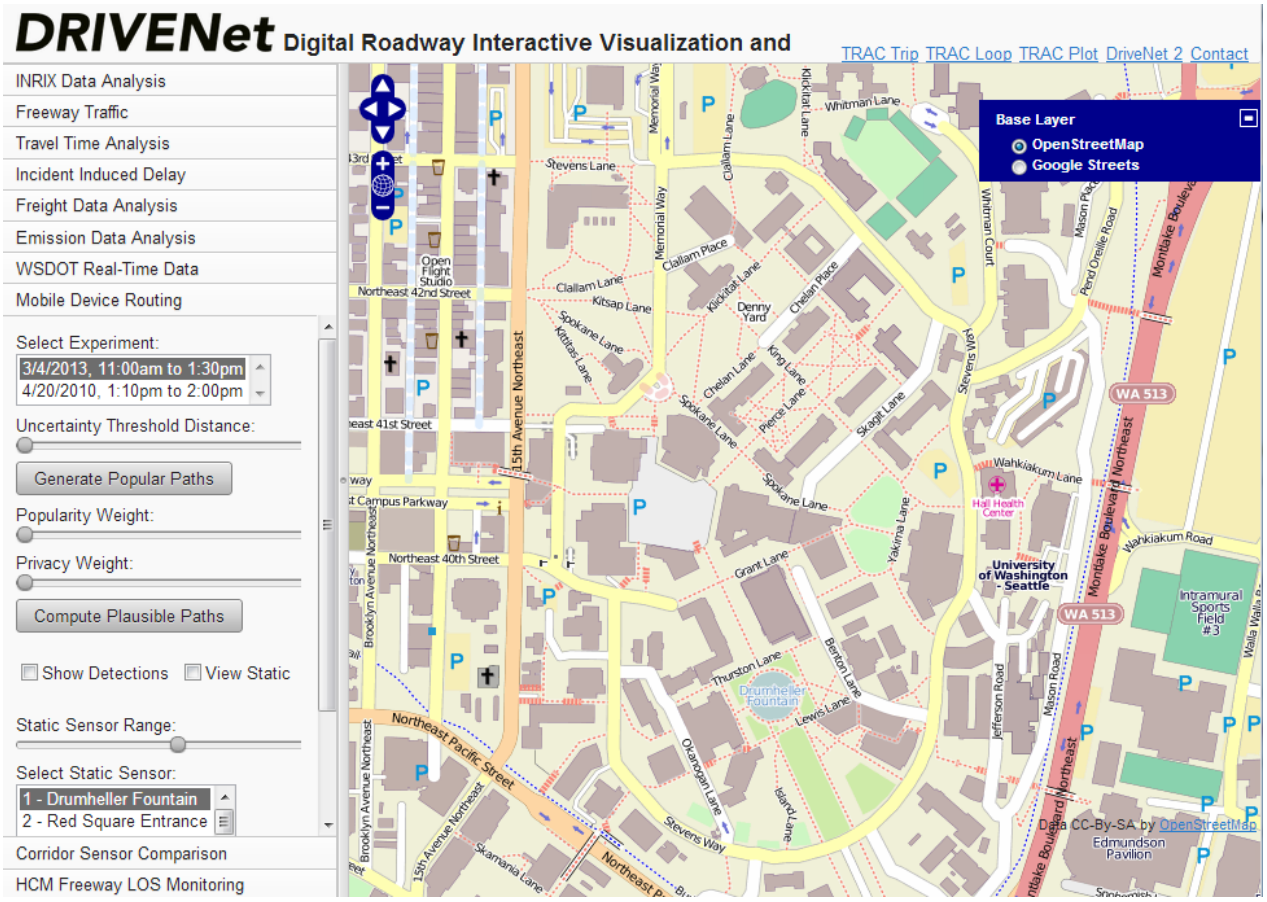


Figure 4-8 High Resolution OpenStreetMap near University of Washington

Figure 4-9 describes how clients dynamically interact with OpenStreetMap in the DRIVENet system and the backend processes. When a web server receives clients' request for a map, it transmits the request to the OSM mapping server for retrieving map contents. The OSM mapping server renders the map with specified geospatial information and sends it back to the web server. Web server then passes map contents to clients. At the client side, OpenLayers provides the service to obtain map images from servers and display map tiles on the screen (Haklay et al., 2008).

OpenLayers is an open-source JavaScript library running at the client side which helps users interact with dynamic maps from disparate services. Extra features are provided by OpenLayers. Specifically, it allows developers to lay numerous data on top of map layers, such as vector layers, markers, and pop-up windows, as *Figure 4-10* demonstrates.

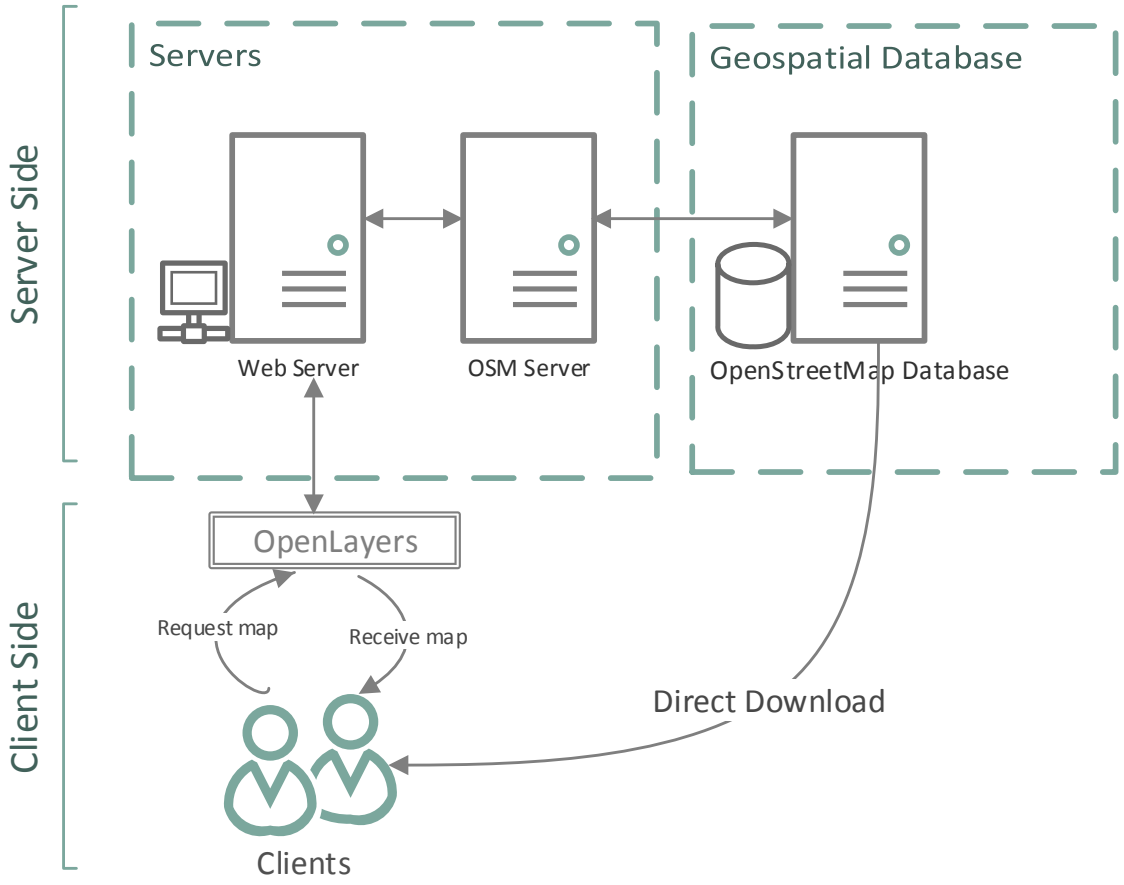


Figure 4-9 How to interact with OpenStreetMap

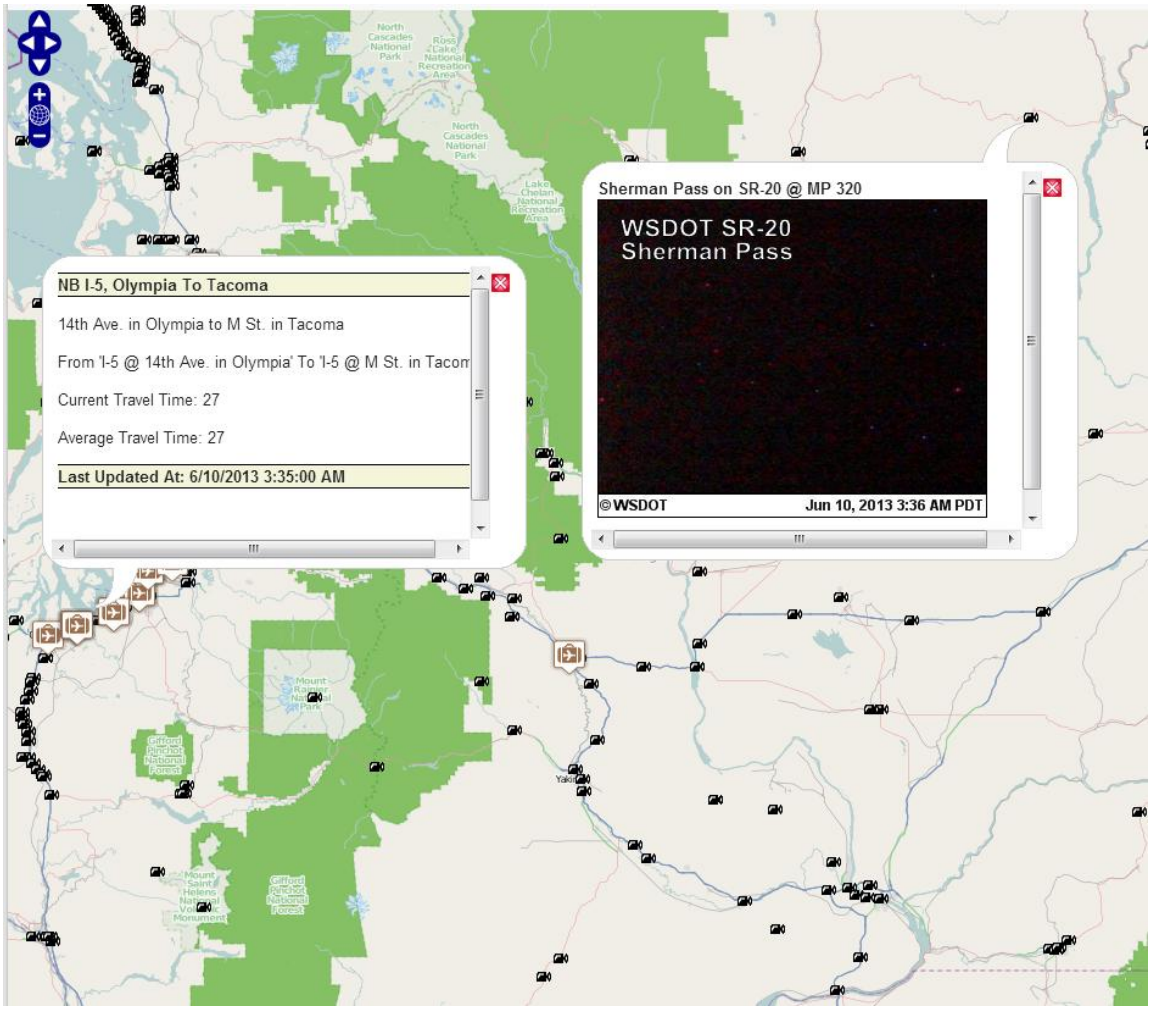


Figure 4-10 Multiple Layers on Top of Map

4.3.2 PostgreSQL, PostGIS, and pgRouting

The biggest challenge in the previous DRIVENet system is the lack of geo-processing power, which makes it lose the capability of spatial modeling. In the new system, PostgreSQL with extender PostGIS and pgRouting is adopted to maintain geo-data and perform spatial modeling, as the relationships outlined in *Figure 4-11*. Those three products are all free, open source, and well-supported by their active communities. Although some commercial software such as ArcGIS/ArcServer could perform same jobs, open source projects are always more academic in

nature despite the fact that commercial products usually have expensive license and usage restrictions. In the rest of this section, more details about PostgreSQL, PostGIS, and pgRouting are introduced.

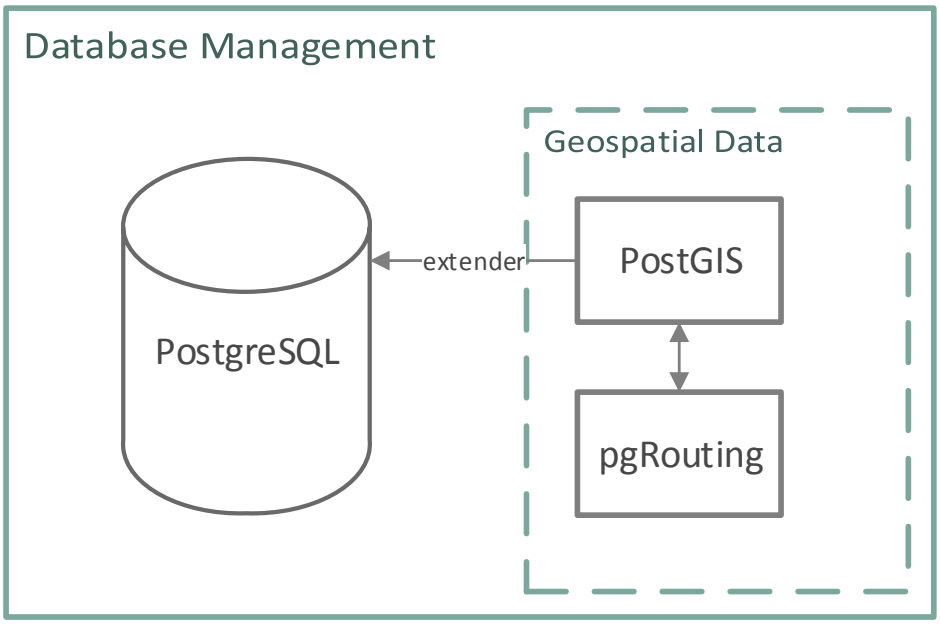


Figure 4-11 PostgreSQL, PostGIS, and pgRouting

PostgreSQL is a sophisticated and feature-rich object-relational database management system under an open source license (PostgreSQL, 2013). Its powerful functions and efficient performance make it the most popular open source database and be able to compete against well-known commercial products, such as Oracle, IBM DB2, and Microsoft SQL server. Some advanced and unique features make it distinguished from others, including table inheritance, support for arrays, multiple-column aggregate functions etc. Moreover, the active global community of developers keep updating PostgreSQL with the latest database technology.

With the capacity of PostgreSQL as a tabular database, PostGIS is a spatial database

extender built on PostgreSQL (Obe, 2011). The PostgreSQL/PostGIS combination offers supports to store, maintain, and manipulate geospatial data, making it one of the best choices for spatial analysis. Besides the geo-data storage extension, PostGIS has nearly 300 geo-processing operators or functions. The ability to analyze geographic data directly in the database by SQL sets distinguishes PostGIS from commercial competitors. For example, the following spatial query creates a polygon buffer with size of 10,000 feet:

Select ST_Buffer(the_geom, 10000) from county_polygon

pgRouting is the extension of PostGIS/PostgreSQL geospatial database, which provides a set of routing-related SQL functions (pgRouting, 2013). Various routing algorithms are supported by pgRouting, including shortest path Dijkstra, shortest path A*, shortest path shooting*, traveling salesperson problems, and driving distance calculation. Meanwhile, its open source feature makes it convenient to develop user-specified algorithm and integrate it into pgRouting. More advanced algorithms such as Multimodal Routing support, Two-Way A*, time-dependent/dynamic shortest path algorithm is going to be included soon.

4.3.3 R and Rserve

R is a free and powerful statistical analysis tool utilized by more than two million people for machine learning, statistical modeling, and data visualizations (R, 2013). With thousands of active contributors from academia, R keeps evolving with the latest efficient and innovative algorithms. Meanwhile, R provides excellent tools for creating graphics, which enable users get better insights via data visualization. Rserve, a TCP/IP server connecting to R, integrates R into the DRIVNet

system so that it takes full advantages of R's statistical computation capability (Rserve, 2013). Several modules in the system use the combination of Rserve and R as the major tool for statistical analysis and data visualization, as *Figure 4-12* and *Figure 4-13* demonstrate. By integrating R and its countless statistical and graphic packages, DRIVENet offer an easy and customizable interface to perform complex analysis and data visualization for users even without any background knowledge of R scripts.

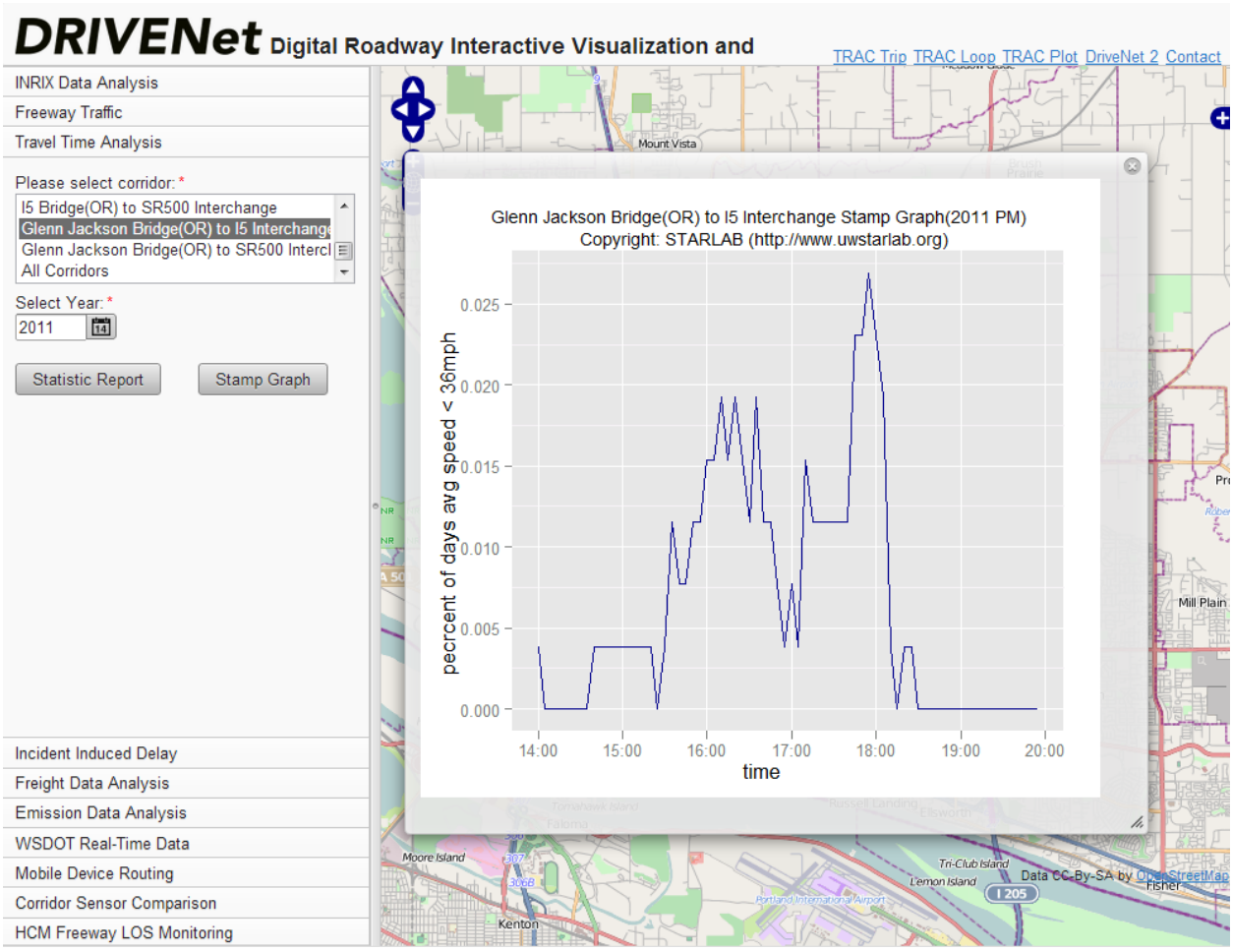


Figure 4-12 Travel Time Performance Measurement

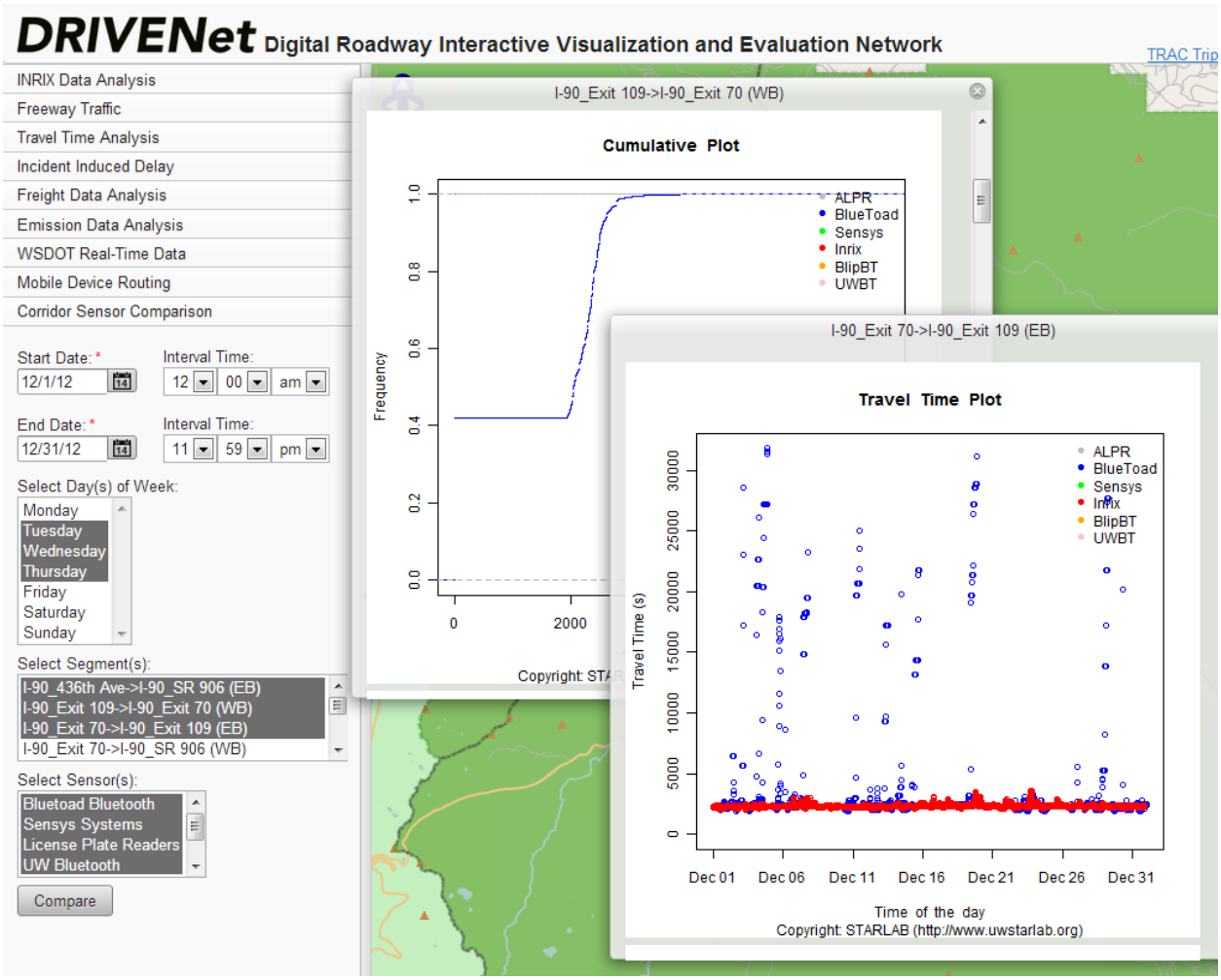


Figure 4-13 Corridor Sensors Comparison

Chapter 5: Real-time Freeway Performance Measurement

To demonstrate the data sharing, integration, visualization, and analysis in the DRIVENet eScience transportation platform, a pilot research on automating network-wide real-time freeway performance measurement is described in Chapter 5 and Chapter 6.

5.1 Background

Real-time freeway performance measurement quantitatively describes traffic conditions to transportation researchers, operators, planners, and the general public in a timely manner. With the network-wide real-time information, decision makers can not only quickly evaluate the quality of service on transportation facilities and identify the congestion bottlenecks, but also perform prompt coordination and may refine policy and investment decisions. The ultimate goal of measuring freeway performance is to improve transportation mobility and accessibility.

The most widely used guidance for measuring freeway performance is the HCM 2010, which has been undergoing constant revision ever since 1944 (Kittelson, 2000). The 2010 version HCM, published by TRB of the National Academies of Science, is a collection of the state-of-the-art methodologies for quantifying the quality of service on transportation facilities. One important concept introduced by HCM is the LOS, which represents a qualitative ranking of traffic performance ranging from **A** to **F**. LOS **A** represents the best traffic operational condition, while **F** is the worst. In this study the HCM 2010 methods are applied for quantifying freeway performance. Although real-time traffic data as well as roadway geometric data are collected by every DOT, there is no universal procedure developed to utilize available datasets and automate

the network-wide freeway operational analysis. FREEVAL 2010, the computational engine executed in Microsoft Excel, is one alternative solution to freeway facilities analysis (HCM, 2010). However, FREEVAL requires users to manually input geometric and traffic demand information for each segment, which can be extremely cumbersome when analyzing long roadway segments across multiple time periods. With the significant computational power and comprehensive data in the data warehouse (such as mainline loop detector data, freeway geometric factors, INRIX speed, etc.), DRIVENet apparently provides a mature platform to perform real-time LOS analysis for freeway segments. Due to the limited information on ramp geometrics, on-ramp volume, off-ramp volume, and weaving volume, this study will only focus on quantifying traffic operational performance for basic freeway segment.

5.2 Challenges

The methodologies in HCM 2010 has limitations. First, HCM methods can only be applied to local oversaturated conditions but not system-wide. Second, some special conditions are not taken into account, such as segment near toll plaza, free-flow speed above 75 mph, or free-flow speed below 55 mph. Although HCM recommends potential alternative tools to fill those gaps, most of them are commercialized simulation tools. Considering the cost and technical challenges, it is not an ideal solution if we perform the real-time analysis in DRIVENet.

Measuring network-wide performance poses challenges on integrating multiple geospatial data layers. Different GIS data layers have different line segments, even when they share the same route, start point, and end point. For example, in *Figure 5-14*, the same route I-5 northbound from milepost 0 to milepost 10 is segmented into different lines in different GIS data layers. One

possible solution is to use line-to-line vector overlay, as *Figure 5-15* shows. However, the operation of network-wide multi-layer overlay on the fly is inefficient and time-consuming. Better spatial data fusion techniques needs to be used for integrating multiple geo-data sources with efficiency and accuracy.

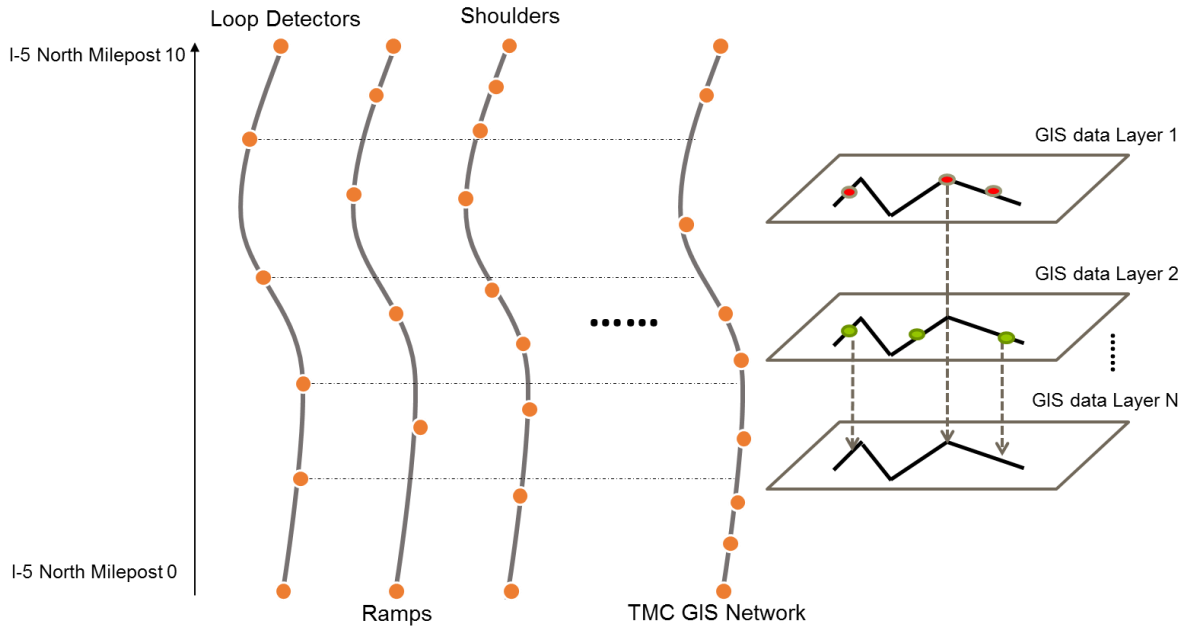


Figure 5-14 Geospatial Data Fusion Challenge

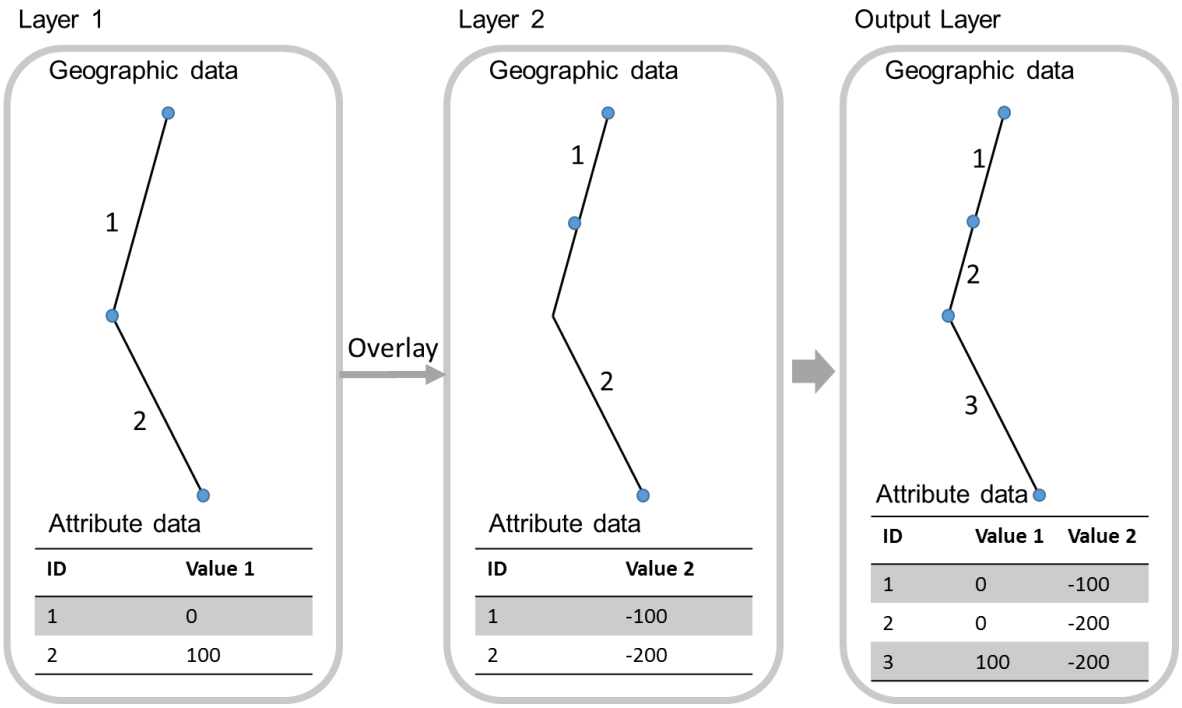


Figure 5-15 Vector Overlay

The objective of this case study is to automate the freeway performance measurement in a consistent, efficient and accurate manner, given existing resources including geometric factors, loop detector data and INRIX speed data. The DRIVENet platform is utilized to implement the automation, not only because of its interoperable data framework but also its customizable computing power. The rest of this chapter elaborates on the spatial modeling framework of network-wide freeway performance measurement.

5.3 Modeling Framework

The modeling process is divided into two main phases as shown in *Figure 5-16*. In the first stage, the roadway network is segmented using an innovative spatial data fusion technique - pixel-based

segmentation. Once the segmented network is formed, three different methods are applied to compute LOS in phase 2, namely, HCM 2010 Method, HCM 2010 Method with INRIX Speed Data, and Multi-regime Prediction Method.

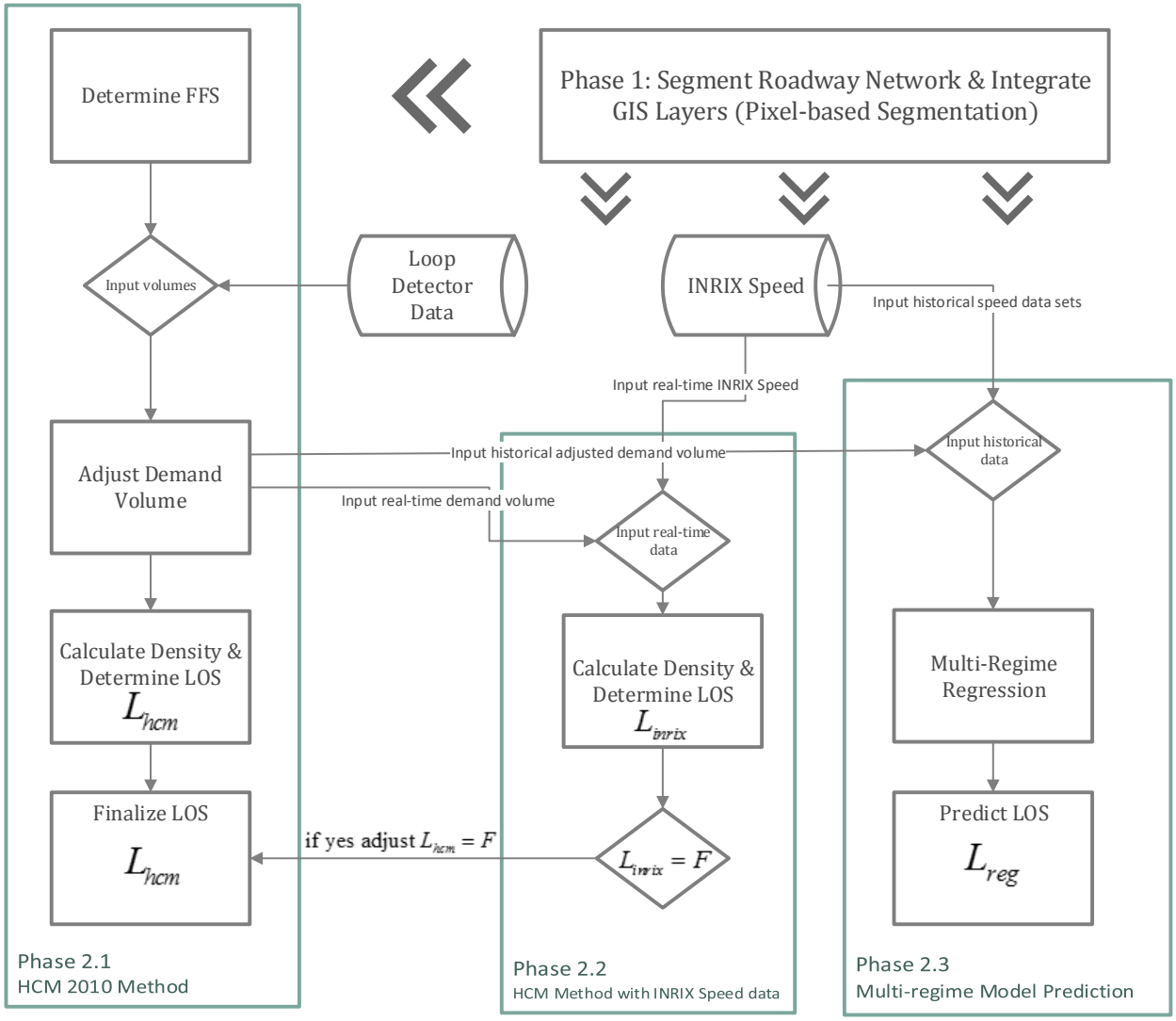


Figure 5-16 Modeling Framework

5.3.1 Segment Roadway Network and Integrate GIS Layers

With heterogeneous datasets, it is inevitable to perform multi-layer geospatial data processing, to superimpose multiple GIS layers to generate an output layer. To calculate performance measurements, a fundamental network layer needs to be prepared, in which each basic roadway segment has the same attribute data as input value. Particularly, the HCM 2010 requires roadway to be segmented uniformly. Uniform segments must share the same attribute data, including geometric features and traffic features. In GIS, vector overlay is the common and major solution to combine both the geographic data and attribute data from multiple input GIS layers as *Figure 5-15* presents. However, in our case, the network-wide large volume spatial data makes the overlay analysis time consuming and computational intensive. Additionally, if a new GIS layer is imported into the DRIVENet data warehouse, it is not realistic to re-perform the entire overlay operations.

Therefore, pixel-based segmentation, a novel method to model the geospatial data, is proposed, which borrows the concept of pixel in digital imaging. A pixel is generally treated as the fundamental unit of a digital photo, extracted from the words "PICture ELement" (Wikipedia, 2013). Millions of pixels are combined together to resemble the original seemingly. The quality of the image highly depends on the total number of pixels used, which is defined as resolution. As *Figure 5-17* indicates, the more pixels the image contains, the more details it is able to reveal.

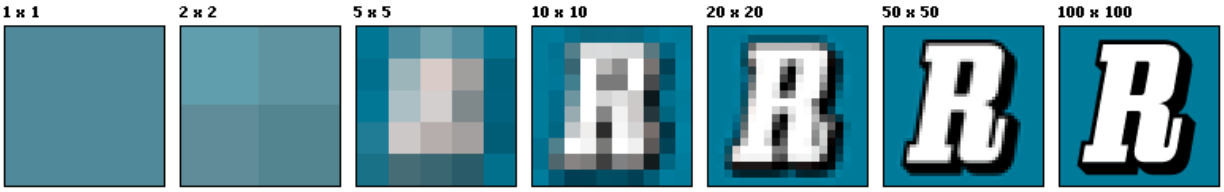


Figure 5-17 Image Resolution (Wikipedia, 2013)

Similarly, pixel-based segmentation subdivides a roadway network into basic segments of equal length, called line pixel. The length of line pixel defines the resolution of segmentation. The shorter the pixel length is, the more details the output network contains. For instance, *Table 5-5* illustrates I-5 northbound with start milepost 140.4 and end milepost 140.9, is subdivided into 5 basic segments of equal length (0.1 mile each). Attribute data of output network use the combination of route ID, start milepost, and end milepost as the unique key to link with the geographic data. With the geographic data being segmented into equal line pixels already, the process of superimposing multiple GIS layers can be accomplished in attribute data side only. As the Linear Referencing System (LRS) WSDOT adopts to identify the locations of features is based on state route ID and feature distance in miles from route beginning (WSDOT's Linear Referencing System, 2013), it is easy and fast to retrieve corresponding feature given the route ID, start milepost and end milepost. Pseudocode for integrating attribute data from multiple GIS layers can be found below:

```
function integrateGISLayers
    for each route r in network
        for k = 0; k < r.length; k = k + pixel_length
            start_mp = k;
            end_mp = k + pixel_length;
            for each input GIS Layers l
                # look up attribute data of l
                # given routeid, start_mp and end_mp
                outputLayer[r, start_mp, end_mp, l]
                    = getAttributeDate(I, r, start_mp, end_mp);
            end
        end
    end
    output outputLayer;
```

Route	Start MP	End MP	Direction	Shoulder width	Rdwy width	NumLns	Avg width	Urban Rural	Terrain	TRD	Upper Ramp MP	Lower Ramp MP
5	140.4	140.5	North	10	48	4	12	U	Level	0.8333	141.64	138.04
5	140.5	140.6	North	10	48	4	12	U	Level	0.8333	141.64	138.04
5	140.6	140.7	North	10	48	4	12	U	Level	1	141.64	138.04
5	140.7	140.8	North	10	48	4	12	U	Level	1.1666	141.64	138.04
5	140.8	140.9	North	10	48	4	12	U	Level	1.1666	141.64	138.04

Table 5-5 Segmented I-5

The pixel-based segmentation is used in this study for the following reasons: First, it separates the attribute data from geographic data. Compared to the vector overlay operations, the integration of attribute data based on LSR is more efficient, fast and easy to implement. Second, the fixed segmentation makes it convenient to integrate more GIS layers into existing network in the future, as long as the pixel resolution remains the same. Third, the value of pixel resolution is flexible for us to decide the level of accuracy to achieve. If the line pixel is infinitely close to 0, the output attribute table will capture perfect details no matter how many GIS layers are imported. In reality, pixel size 0.1 mile is a good choice to balance the efficiency and accuracy.

5.3.2 Calculate LOS using the HCM 2010 methodology

Due to limitation of available datasets, this study will only focus on the LOS calculation for basic freeway segments. The HCM 2010 provides a comprehensive method for analyzing the LOS as demonstrated in *Figure 5-16 Phase 2.1*. Notice that there is no measured FFS available for the entire network layer, FFS is computed by lane width adjustment and lateral clearance adjustment in this study. The HCM 2010 is unable to handle system-wide oversaturated flow conditions, and only focuses on analyzing under-saturated flow conditions. Over-saturated flow conditions are discussed in the next section.

Step 1: Input Data

In this step, demand volume, number and width of lanes, right-side lateral clearance, total ramp density, percent of heavy vehicles, peak hour factor, terrain, and the drive population factors are retrieved from the DRIVENet data warehouse.

Demand Volume

Real-time demand volume are mainly estimated from loop detectors. The system automatically fetches all the cabinets between the Nearest Upstream Ramp (NUR) and the Nearest Downstream Ramp (NDR), and then queries corresponding latest 15-min flow. Demand volume is calculated using the following equation:

Equation 5-1

$$V = 4 \times median(\{latest\ 15minute\ flow\ |cabinest\ between\ NUR\ and\ NDR\})$$

$$V: hourly\ volume\ (veh/h)$$

Median is selected to measure the central tendency since it naturally eliminate the outliers. It is then multiplied by 4, which projects into hourly volume. For instance, in *Figure 5-18*, there are a total of six cabinets between upstream and downstream ramps. The 15-min flows fetched are shown as 500, 100, 450, 450, and 550. Hence, hourly volume for the segments between upstream and downstream ramps equals to 450×4=1800 *veh/h*. Notice that if there are no cabinets/loop detectors between the upstream and downstream ramps, the system will mark there is no demand volume input for segments and it will use real-time INRIX speed and historical regression model

to predict LOS, which will be introduced later in this chapter.

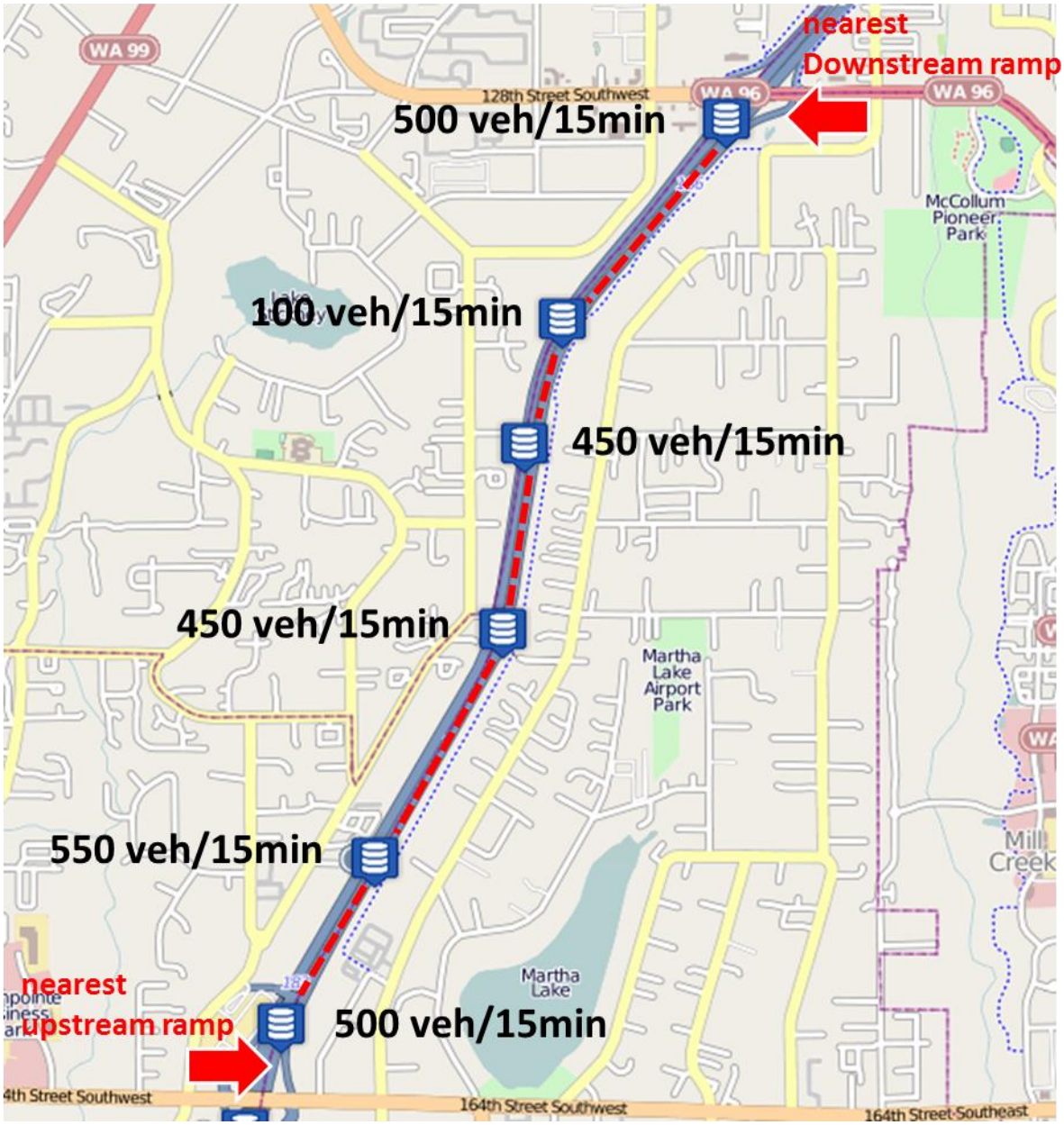


Figure 5-18 Nearest upstream and Downstream Ramps

Total Ramp Density (TRD)

Total ramp density is defined as the total number of ramps (both on and off with one direction) within 2miles of midpoint of segment under study. Given the study segment start milepost and end milepost, the following equation could be used to calculate TRD:

Equation 5-2

$$midpoint = \frac{start\ milepost + end\ milepost}{2}$$

$$TRD = total\ number(\{ramps \mid ramp\ route = segment\ route\ \boldsymbol{and}\ ramp\ milpost\ beween\ (midepoint + 3) and\ (midpoint - 3)\ \})/\ 6\ miles$$

Other Input Data

The geometric data, including number and width of lanes, right-side lateral clearance and terrain, are originally downloaded from WSDOT Roadway Datamart for GIS. Geospatial data fusion has performed using the methods introduced in the previous section. Since there is no site-specific data available for the remaining features, default values recommended by NCHRP Report 599 (Zegeer et al., 2008) are adopted.

Required Data	Default Values
Peak Hour Factor	Urban: 0.92, Rural: 0.88
Driver Population Factor	Urban: 1.0, Rural: 0.975
Percentage of heavy vehicles (%)	Urban: 5%, Rural: 12%

Table 5-6 Default Values for Basic Freeway Segments

Step 2: Determine Free-Flow Speed

Since the site-specific measured FFS is not available, the following equation developed by HCM 2010 is used to estimate FFS. Lane width, right-shoulder lateral clearance, and ramp density are taken into account to adjust the Base Free-Flow Speed (BFFS). The estimated FFS is then rounded to the nearest 5 mph as HCM suggests. The adjustment value can be found in HCM 2010.

Equation 5-3

$$FFS = 75.4 - f_{LW} - f_{LC} - 3.22\,TRD^{0.84}$$

Where

$FFS = estimated\ free\ flow\ speed\ in\ mph$

$f_{LW} = lane\ width\ adjustment\ in\ mph$

$f_{LC} = lateral\ clearance\ adjustment\ in\ mph$

$TRD = total\ ramp\ density\ adjustment\ in\ mph$

Step 3: Adjust Demand Volume

Demand volume obtained from loop detectors needs to be converted into service flow rate under equivalent base conditions. According to the HCM 2010, the base conditions for a basic freeway segment are specified as

- 12-ft lane widths
- 6-ft right shoulder clearance
- 100% passenger cars in the traffic stream

- Level terrain
- A driver population of regular users familiar with roadway in general

Equation 5-4 below is then utilized for the conversion:

Equation 5-4

$$v_p = \frac{V}{PHF \times N \times f_{HV} \times f_p}$$

Where

$v_p = adjusted\ demand\ volume\ under\ base\ conditions\ in\ pc/h/ln$

$V = hourly\ demand\ volume\ under\ prevaling\ conditions\ in\ veh/h$

$PHF = peak\ hour\ factor$

$N = numer\ of\ lanes$

$f_{HV} = heavy\ vehicle\ adjustment\ factor$

$f_p = driver\ population\ adjustment\ factor$

The heavy-vehicle adjustment factor could be calculated by the following equation

Equation 5-5

$$f_{HV} = \frac{1}{1 + P_T(E_T - 1) + P_R(E_R - 1)}$$

Where

$f_{HV} = heavy\ vehicle\ adjustment\ factor$

$P_T = percentage\ of\ trucks\ and\ bus\ in\ the\ traffic\ stream$

$P_R = percentage\ of\ recreational\ vehicles\ in\ the\ traffic\ stream$

$E_T = passenger\ car\ equivalent\ factor\ for\ trucks\ and\ bus$

$E_R = passenger\ car\ equivalent\ factor\ for\ recreational\ vehicles$

As HCM suggests, the proportion of recreational vehicles in the traffic stream is small and close to 0 in many cases. Hence, in this study, P_R is set to be 0 as the default value. The value of passenger car equivalent factors E_T and E_R are also recommended by HCM 2010 based on type of terrain or grades.

Step 4: Calculate Density and Determine LOS

Given the FFS from **Step 2** and adjusted volume v_p from **Step 3**, the average passenger car speed S can be found in *Figure 5-19* or computed by speed-flow equation in *Table 5-7*. Then the density D_{hcm} can be derived:

Equation 5-6

$$D_{hcm} = \frac{v_p}{S}$$

Once the density is computed, the LOS L_{hcm} can be determined from *Table 5-8*.

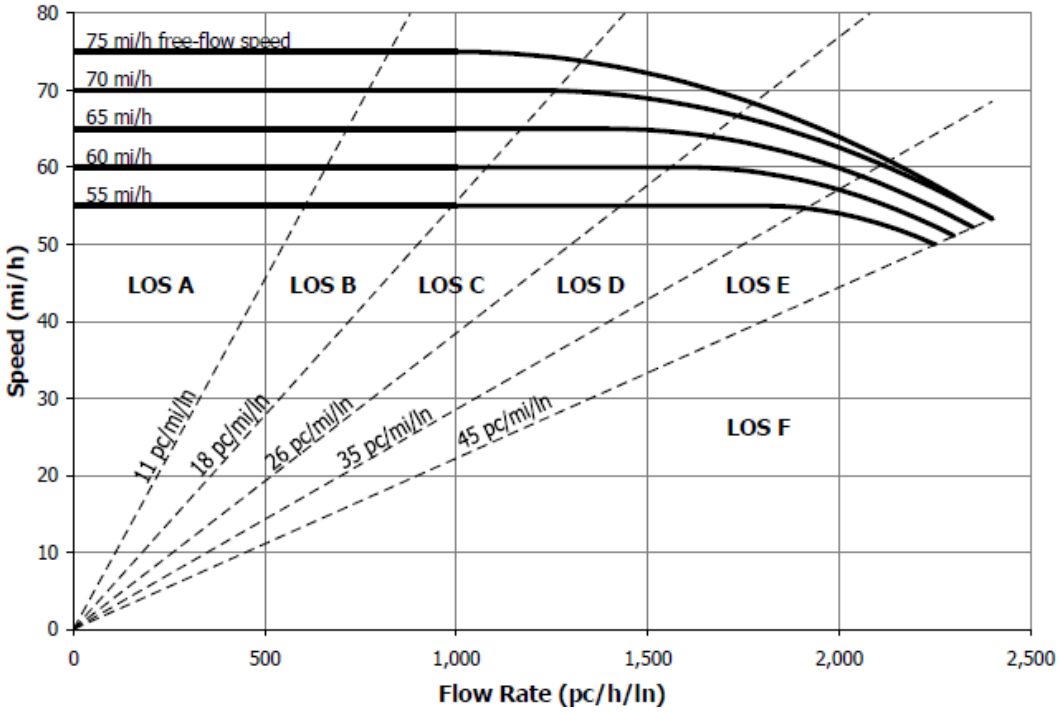


Figure 5-19 HCM Speed-Flow Model (HCM, 2010)

FFS (mi/h)	Break-Point (pc/h/ln)	Flow Rate Range	
		≥0 ≤ Break-Point	>Break-Point ≤ Capacity
75	1,000	75	$75 - 0.00001107\ (v_p - 1{,}000)^2$
70	1,200	70	$70 - 0.00001160\ (v_p - 1{,}200)^2$
65	1,400	65	$65 - 0.00001418\ (v_p - 1{,}400)^2$
60	1,600	60	$60 - 0.00001816\ (v_p - 1{,}600)^2$
55	1,800	55	$55 - 0.00002469\ (v_p - 1{,}800)^2$

Notes: FFS = free-flow speed, v_p = demand flow rate (pc/h/ln) under equivalent base conditions.
Maximum flow rate for the equations is capacity: 2,400 pc/h/ln for 70- and 75-mph FFS; 2,350 pc/h/ln for 65-mph FFS; 2,300 pc/h/ln for 60-mph FFS; and 2,250 pc/h/ln for 55-mph FFS.

Table 5-7 Speed-Flow Equations (HCM, 2010)

Density	LOS
11	A
18	B
26	C
35	D
45	E

Table 5-8 LOS Criteria for Basic Freeway Segments

5.3.3 Incorporate the real-time INRIX speed into LOS calculation

One of the limitations of the HCM method is that it cannot analyze system-wide oversaturated conditions. In other words, once demand is greater than capacity, HCM is unable to estimate space mean speed as well as density. However, in reality, it is critical to identify oversaturated conditions spatially and temporally so that operators and planners can understand bottleneck (formation, propagation, and dissipation) of the facilities. As suggested by *Figure 5-20*, under oversaturated conditions, the traffic speed drops dramatically, typically below 35 mph. To fill the gap of analyzing oversaturated conditions, INRIX speed data is utilized in the study and incorporated in the LOS calculation. With the demand volume still obtained from loop detectors and adjusted by the HCM 2010 methodology, INRIX speed S_{inrix} is utilized to estimate the density as shown in *Equation 5-7*:

Equation 5-7

$$D_{inrix} = {}^{v_p}/_{S_{inrix}}$$

$$\delta = \begin{cases} 1 \; if \; D_{inrix} \leq 45 \\ 0 \; if \; D_{inrix} > 45 \end{cases}$$

$$D_{hcm} = \delta \cdot D_{hcm} + (1 - \delta) \cdot D_{inrix}$$

where

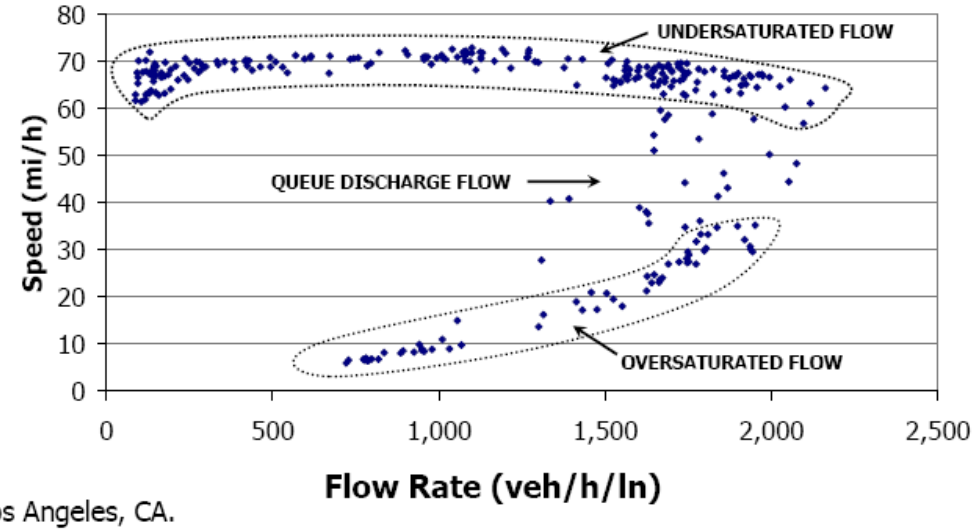
$D_{inrix} = density \; calcuated \; by \; INRIX \; speed$

$D_{hcm} = density \; calcuated \; by \; HCM \; methods$

$v_p = adjusted \; demand \; volume \; under \; base \; conditions \; in \; pc/h/ln$

$S_{inrix} = real \; time \; INRIX \; speed$

Additionally, using INRIX speed to estimate LOS also provides ground-truth data to validate the feasibility of HCM methodologies, which is discussed in Chapter 6.



Note: I-405, Los Angeles, CA.
Source: Caltrans, 2008.

Figure 5-20 Undersaturated, Queue Discharge, and Oversaturated Flow (HCM, 2010)

5.3.4 Develop Empirical Speed-Density Regression Equations to Predict LOS

As one of the primary concerns, the quality of traffic data greatly influences the accuracy in performance estimation. The data quality issues involve at least (1) missing data, (2) suspicious or erroneous data, and (3) inaccurate data (Turner, 2001). While erroneous data do not follow accepted principles or go beyond thresholds, inaccurate data contains inexact values due to measurement error. In this study, these three types of errors are all treated as the invalid traffic data entry. The data quality issues trigger two major challenges: (1) How to identify the bad data? (2)

How to compensate for the invalid data input?

Much efforts have been made to develop comprehensive and sophisticated quality checking methods. In practice, threshold approach is often adopted to ensure the sensor value fall within a reasonable range. The combination of volumes, speed, and occupancies provides relatively straightforward yet robust way to check data error. Jacobson *et al*. developed an algorithm which uses volume-to-occupancy ratios to examine the reliability of loop detector data (1990). In addition, time series of traffic samples can be used for comparison. For example, Chen *et al.* (2003) proposed a diagnostics algorithm to efficiently find malfunctioning single-loop detectors based on sequence of volume and occupancy measurement for the entire day. Ishak (2003) developed a fuzzy-clustering approach to measure the uncertainties of freeway loop detector. Moreover, spatial relationship between detectors also turns out to be an effective tool to accurately detect errors. Kwon *et al.* (2004), for instance, utilized the strong measurement correlations between upstream and downstream sensors to detect spatial configuration errors.

All those advanced algorithms demonstrate robust solutions in identifying the quality issue of loop detectors. It thus leads to another question on how to estimate real-time density or LOS when the input demand volume is invalid. With the relatively comprehensive speed dataset from INRIX, this research focuses on predicting real-time density given the historical traffic data and real-time speed, as the solution to deal with invalid input volume.

Empirical speed-density relationships provides the most abundant source to perform predictions. Over the past few decades, much research has been done on developing speed-density model. Considering the data-driven nature, multi-regime model based on cluster analysis (Sun *et*

al., 2005) is adopted to fit empirical speed-density observations. This method first applies *K*-means algorithm to traffic datasets, which naturally partitions the data into homogenous groups. It then applies a series single-regime models to find out the one that best fits the data, such that breakpoints can be automatically determined. Notice that Sun's method chooses *k* value by trial-and-error, in this study, the optimal number of clusters are determined by the average Silhouette criterion instead of trial-and-error. For conceptual testing purpose, only linear, logarithmic, and exponential models are included. Pseudocode for building multi-regime traffic model can be found below:

```
function PerformSpeedDensityRegression
    # Given traffic datasets observations
    # Choosing k using the Sihouette
    k = DetermineKbySihouette(observations);
    clusters = kmeans(observations, k);
    for each cluster c in clusters
        # three basic functions chosen to fit c
        lmReg = lm(c.speed ~ c.density, data = c);
        logReg = lm(c.speed ~ ln(c.density), data = c);
        expReg = lm(c.speed ~ exp(c.density), data = c);

        #choose the regression model fits best
        bestReg = max(lmReg.Rsquare, logReg.Rsqaure, expReg.Rsquare);
        output bestReg;
    end
```

Chapter 6: Implementation Results

The aforementioned modeling framework is implemented in a real-world network for pilot testing purposes. I-5 Northbound corridor in Seattle, Washington from milepost 140 to milepost 195 is selected as the study site. It is the primary travel route connecting Tacoma-Everett through Downtown Seattle and has the most comprehensive traffic data available. *Figure 6-21* shows cabinets deployed by WSDOT along the corridor totaled 140. In the next several subsections, network segmentation and data preprocessing is briefly introduced, followed by an elaboration on LOS results computed from three proposed methods, namely, the HCM 2010 method, the HCM 2010 method with INRIX speed, and the multi-regime regression method. The satisfactory results further confirms the reliability and feasibility of proposed modeling framework.

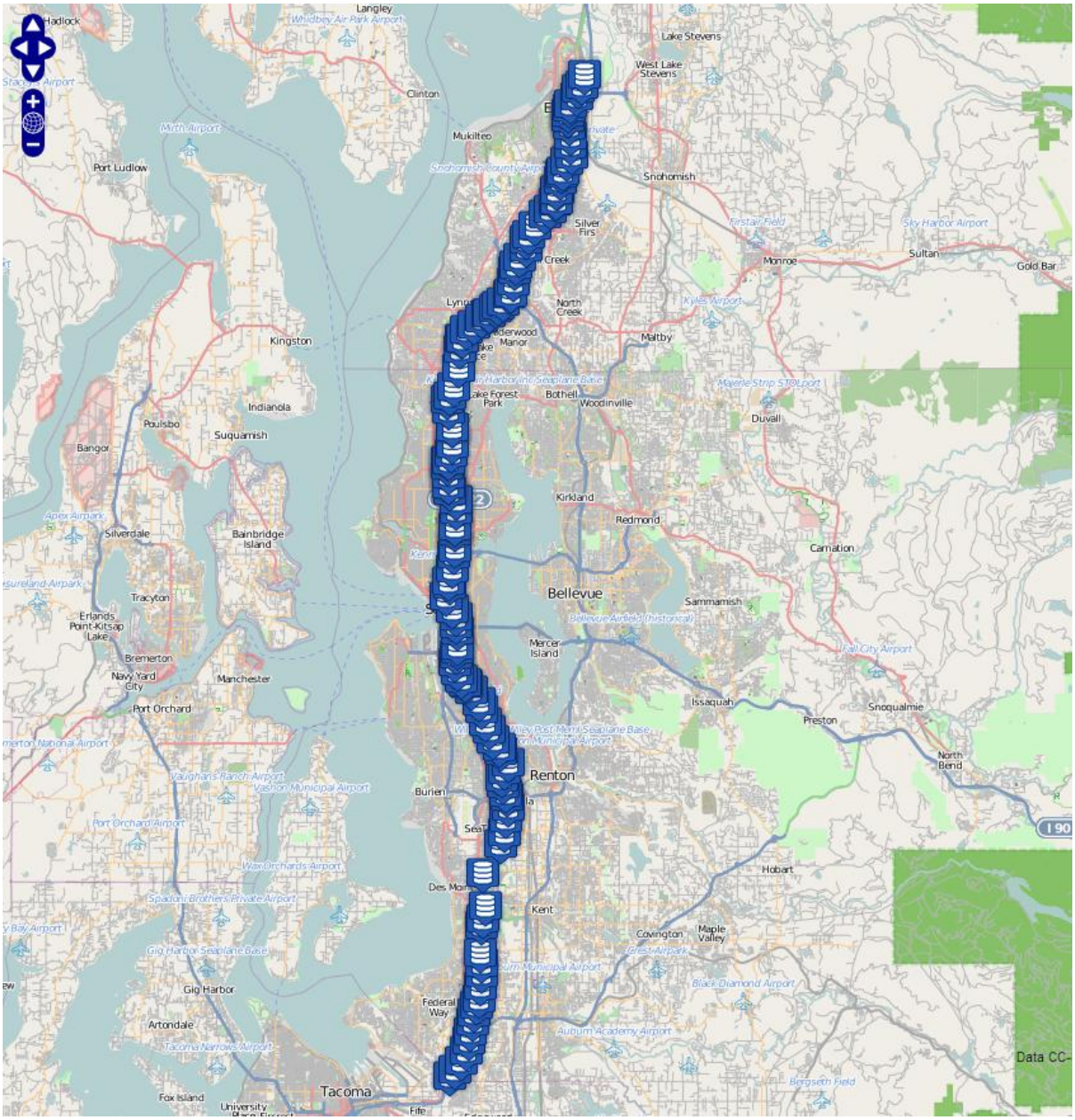


Figure 6-21 I-5 Northbound Corridor (Tacoma - Everett)

6.1 Network Segmentation

Applying pixel-based segmentation on geographic data introduced in Chapter 5, the corridor is subdivided into 550 basic freeway segments with pixel length 0.1 mile. The corresponding

attribute data are then fused according to route ID (I-5), start milepost, and end milepost. *Table 6-9* presents the sample attribute data. Notice that roadway geometric data is relatively static and not updated very often. It is more efficient and effective to pre-process the attribute data fusion instead of running it on the fly.

Route	Start MP	End MP	Direction	Shoulder width	Rdwy width	NumLns	Avg width	Urban Rural	Terrain	TRD	Upper Ramp MP	Lower Ramp MP
5	140.4	140.5	North	10	48	4	12	U	Level	0.8333	141.64	138.04
5	140.5	140.6	North	10	48	4	12	U	Level	0.8333	141.64	138.04
5	140.6	140.7	North	10	48	4	12	U	Level	1	141.64	138.04
5	140.7	140.8	North	10	48	4	12	U	Level	1.1666	141.64	138.04
5	140.8	140.9	North	10	48	4	12	U	Level	1.1666	141.64	138.04

Table 6-9 Fused Attribute Data

6.2 Volume and Speed Data Sets

Real-time volume data are collected from single loop detectors every 20 seconds and INRIX speed is aggregated every 1 minute based on GPS data, respectively. Both datasets are archived in the DRIVENet database. For the pilot testing purpose, 2-day observations are extracted and utilized in the later computation. The two traffic datasets are further aggregated into 15-min time interval as recommended by HCM. Data quality control techniques are applied to ensure data accuracy. For example, several thresholds are set to eliminate obvious outliers. Comprehensive data quality control is critical to the DRIVENet system. For more detail, please refer to (Wang et. al., 2009).

Figure 6-22 shows the scatter plot of adjusted volume v_p vs. speed S_{inrix} as well as density D_{inrix} vs. speed S_{inrix} for a total of 95,040 observations. Notice that the service volume v_p used in Figure *6-22* is under base conditions, converted from real-time traffic counts following the HCM 2010 methods.

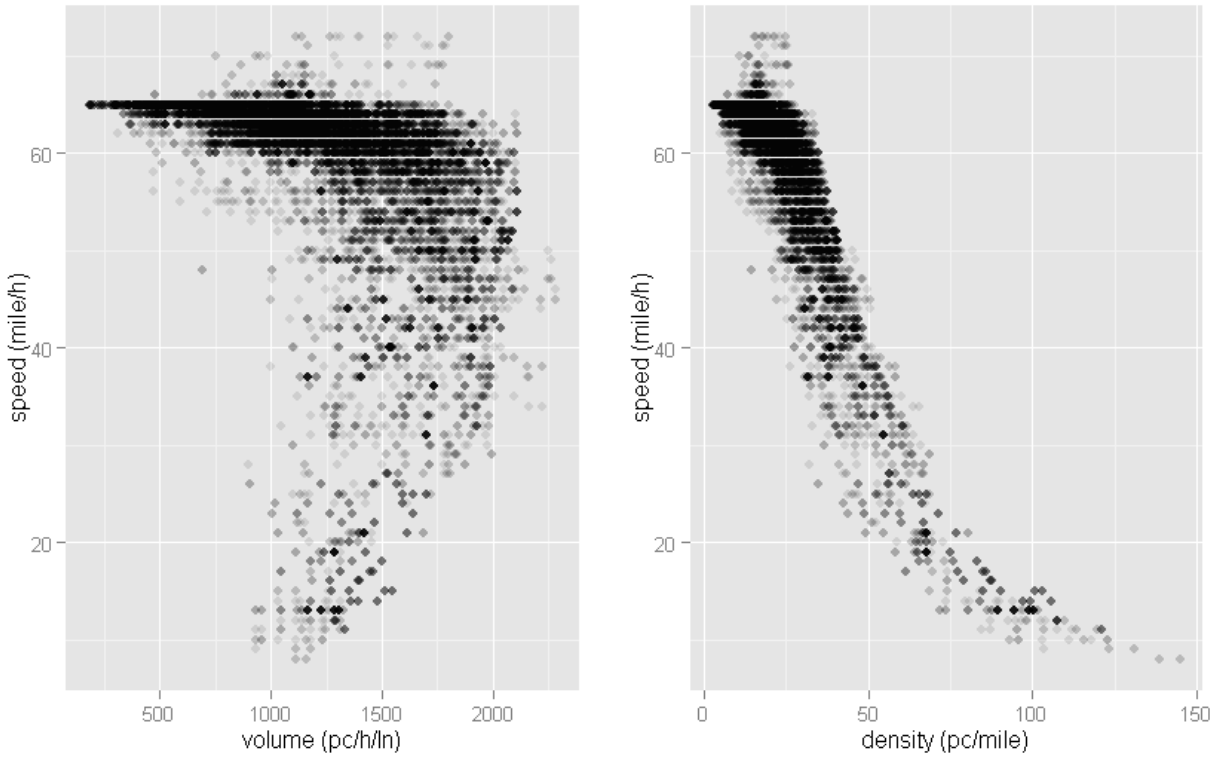


Figure 6-22 INRIX Speed, Adjusted Volume, and Density

6.3 HCM Method with/without INRX Speed Data

The HCM method with volume only and HCM method with volume and speed are applied to compute L_{hcm} and L_{inrix} respectively. Since HCM method is unable to analyze oversaturated conditions (LOS = **F**), the comparison between L_{hcm} and L_{inrix} is conducted for undersaturated flow only. With a total of 92,400 observations fall into the undersaturated conditions, 83.83% of L_{hcm} is equivalent to L_{inrix}, which totaled 77,458 data points. The match rate increases to 98.98% if adjacent LOSs are treated as approximately equal (e.g. *LOS **A**≅LOS **B***). The fact that these two methods have a high consistency in estimating LOS delivers several important messages: (1) the proposed methodologies such as pixel-based segmentation can generate satisfying accuracy; (2)

Using INRIX speed to determine oversaturated condition is feasible and cost-effective; and (3) the quality of INRIX speed data has been justified to some extent, considering the consistency with those results computed by HCM methods in *Phase 2.1 (without INRIX speed)* and *Phase 2.2 (with INRIX speed)*.

Table 6-10 and *Figure 6-23* show the comparison of the LOS category counts produced by the two methods. Note that LOS computed by using INRIX speed usually underestimate service quality. The results are consistent with recent research on transportation sensor comparison conducted by Dr. Yegor Malinovskiy from University of Washington, who found that INRIX speed data usually has smaller standard deviation and underestimate traffic conditions.

LOS	HCM Method	HCM Method with INRIX Speed
A	37430	35994
B	30343	25188
C	18677	20324
D	5756	8077
E	194	2817
F	2640	2640

Table 6-10 LOC Count by Phase 2.1(without INRIX Speed) and Phase 2.2(with INRIX Speed)

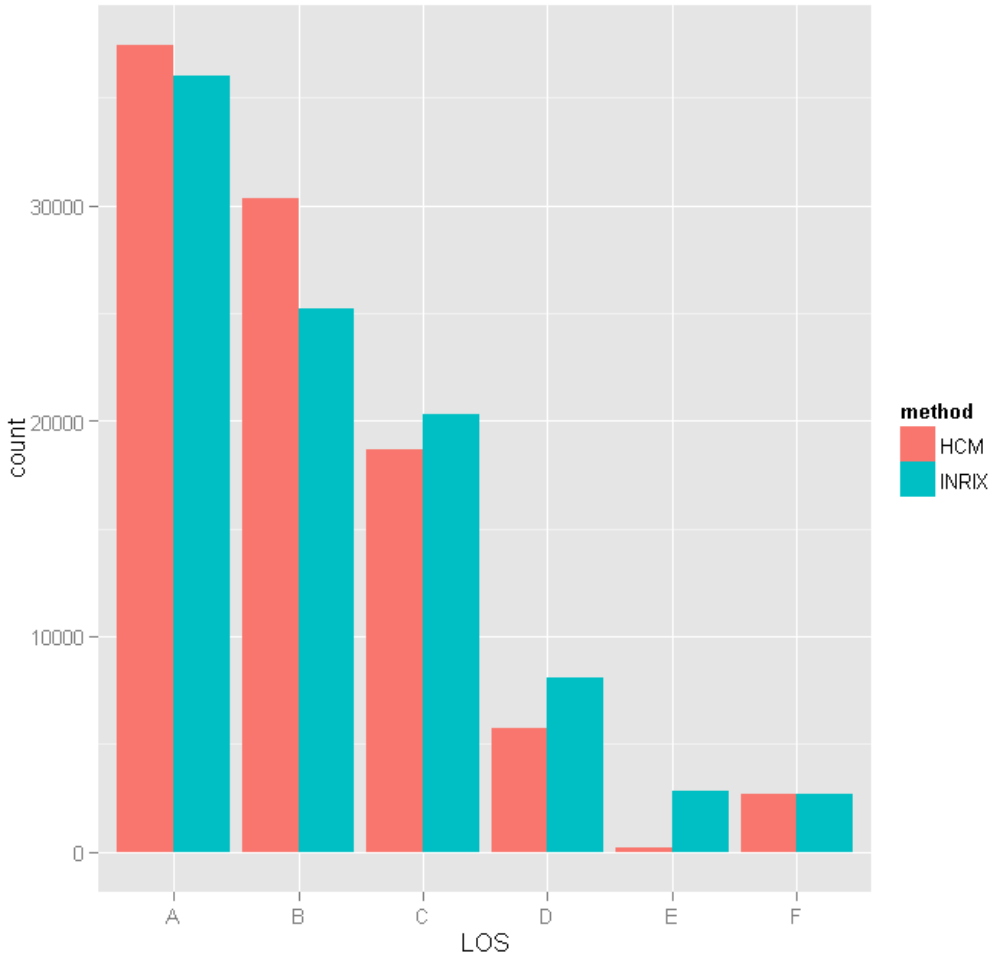


Figure 6-23 LOS by Phase 2.1(without INRIX Speed) and Phase 2.2(with INRIX Speed)

6.4 Regression Analysis

To compensate for missing data or those of low quality, empirical multi-regime density-speed model is used to predict density in this study. During the implementation, the two-day datasets are then divided evenly into training set (November 07 2011) and testing set (November 08 2011) to avoid the overfitting problem.

```
function PerformSpeedDensityRegression
    # Given traffic datasets observations
    # Choosing K using the SilHouette
    K = DetermineKbySilhouette(observations);
    clusters = kmeans(observations, K);
    for each cluster c in clusters
        # three basic functions chosen to fit c
        lmReg = lm(c.speed ~ c.density, data = c);
        logReg = lm(c.speed ~ ln(c.density), data = c);
        expReg = lm(c.speed ~ exp(c.density), data = c);

        #choose the regression model fits best
        bestReg = max(lmReg.Rsquare, logReg.Rsqaure, expReg.Rsquare);
        output bestReg;
    end
```

Following the procedures described in the pseudocode above, *K* value is chosen to be 2 using the Silhouette. According to suggestions from Sun *et al.* (2005), using the original data for *K*-mean algorithm would outperform the normalized data. Hence, this study applies *K*-mean algorithm to the training set without normalization. Clustering results can be found in *Figure 6-24* and *Table 6-11*. As expected, Cluster 1 has high speed and low density which represents free-flow regime, while Cluster 2 has lower speed and high density which represents congested-flow regime.

Three single-regime models, namely, linear, logarithmic, and exponential functions, are then used to fit Cluster 1 and Cluster 2 respectively. The one with the greatest *R squared value* is chosen to represent the empirical speed-density relationship. The following equation shows the final two-regime model obtained from training set:

Equation 6-1

$$u = \begin{cases} 66.3237 - 0.1851k \; if \; k \leq 24.6 \\ \exp(4.657 - 0.02169k) \; if \; k > 24.6 \end{cases}$$

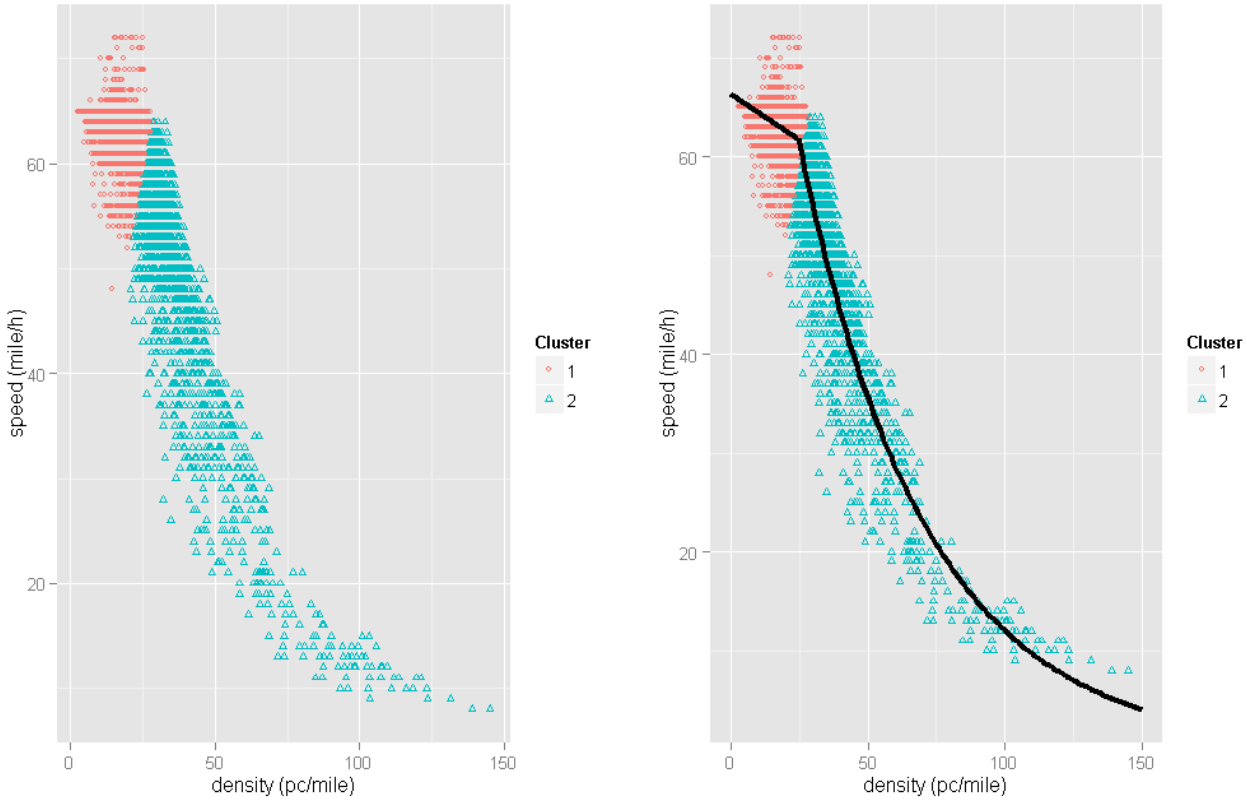


Figure 6-24 Training Set: Two Clusters by K-means Algorithm Analysis

Traffic Dataset		Cluster 1 Center	Cluster 2 Center
I-5 Northbound	Speed (mile/h)	63	53
	Density (pc/mile)	16.94186	32.87736
	Percentage	80.27%	19.73%

Table 6-11 Training Set: Clustering Centers by K-means Algorithm

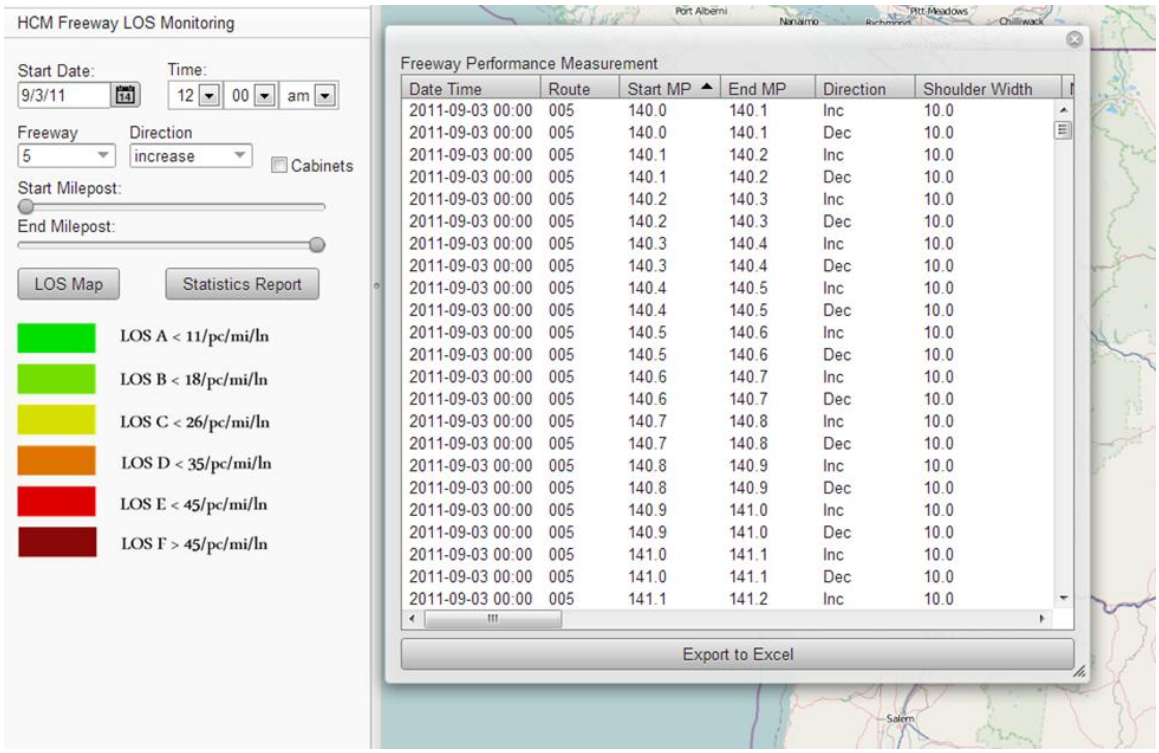
As *Figure 6-24* demonstrates, the two-regime model fits the training set quite well. Comparison between the ground-truth value L_{inrix} and predicted value L_{reg} for both training set and testing set is further conducted. The testing set yields an even lower error as indicated in *Table 6-12*. If adjacent levels are treated as approximately equal, both training error and test error are less than 5% (shown in Accuracy of ± 1 in *Table 6-12*). It thus proves the feasibility and accuracy of the modeling framework proposed in Chapter 5.

Date Set	Accuracy	Accuracy of ±1
Training Set	57.7%	95.38%
Test Set	59.84%	95.01%

Table 6-12 Test Results

6.5 User Interface Design and Data Visualization

Figure 6-25 demonstrates the user interface designed for freeway performance measurement module. The control panel is located on the left side, while interactive map is on the right. Users are free to input date, time, route ID, route direction, start milepost, and end milepost and query the corresponding LOS map by clicking button "LOS Map". As long as the system receives the user request, it will visualize LOS map based on criteria described in color legend on the left. As *Figure 6-26* shows, the LOS map gives a straightforward way to demonstrate LOS spatially, which enables users identify the bottleneck easier. Additionally, related statistics report would be prepared and automatically popped up for downloading if users click the button "Statistics Report". The reports includes detailed information such as segments, geometric factors, speed, density, and LOS, which enables users further analyze the data.



Date Time	Route	Start MP	End MP	Direction	Shoulder Width
2011-09-03 00:00	005	140.0	140.1	Inc	10.0
2011-09-03 00:00	005	140.0	140.1	Dec	10.0
2011-09-03 00:00	005	140.1	140.2	Inc	10.0
2011-09-03 00:00	005	140.1	140.2	Dec	10.0
2011-09-03 00:00	005	140.2	140.3	Inc	10.0
2011-09-03 00:00	005	140.2	140.3	Dec	10.0
2011-09-03 00:00	005	140.3	140.4	Inc	10.0
2011-09-03 00:00	005	140.3	140.4	Dec	10.0
2011-09-03 00:00	005	140.4	140.5	Inc	10.0
2011-09-03 00:00	005	140.4	140.5	Dec	10.0
2011-09-03 00:00	005	140.5	140.6	Inc	10.0
2011-09-03 00:00	005	140.5	140.6	Dec	10.0
2011-09-03 00:00	005	140.6	140.7	Inc	10.0
2011-09-03 00:00	005	140.6	140.7	Dec	10.0
2011-09-03 00:00	005	140.7	140.8	Inc	10.0
2011-09-03 00:00	005	140.7	140.8	Dec	10.0
2011-09-03 00:00	005	140.8	140.9	Inc	10.0
2011-09-03 00:00	005	140.8	140.9	Dec	10.0
2011-09-03 00:00	005	140.9	141.0	Inc	10.0
2011-09-03 00:00	005	140.9	141.0	Dec	10.0
2011-09-03 00:00	005	141.0	141.1	Inc	10.0
2011-09-03 00:00	005	141.0	141.1	Dec	10.0
2011-09-03 00:00	005	141.1	141.2	Inc	10.0

Figure 6-25 User Interface Design

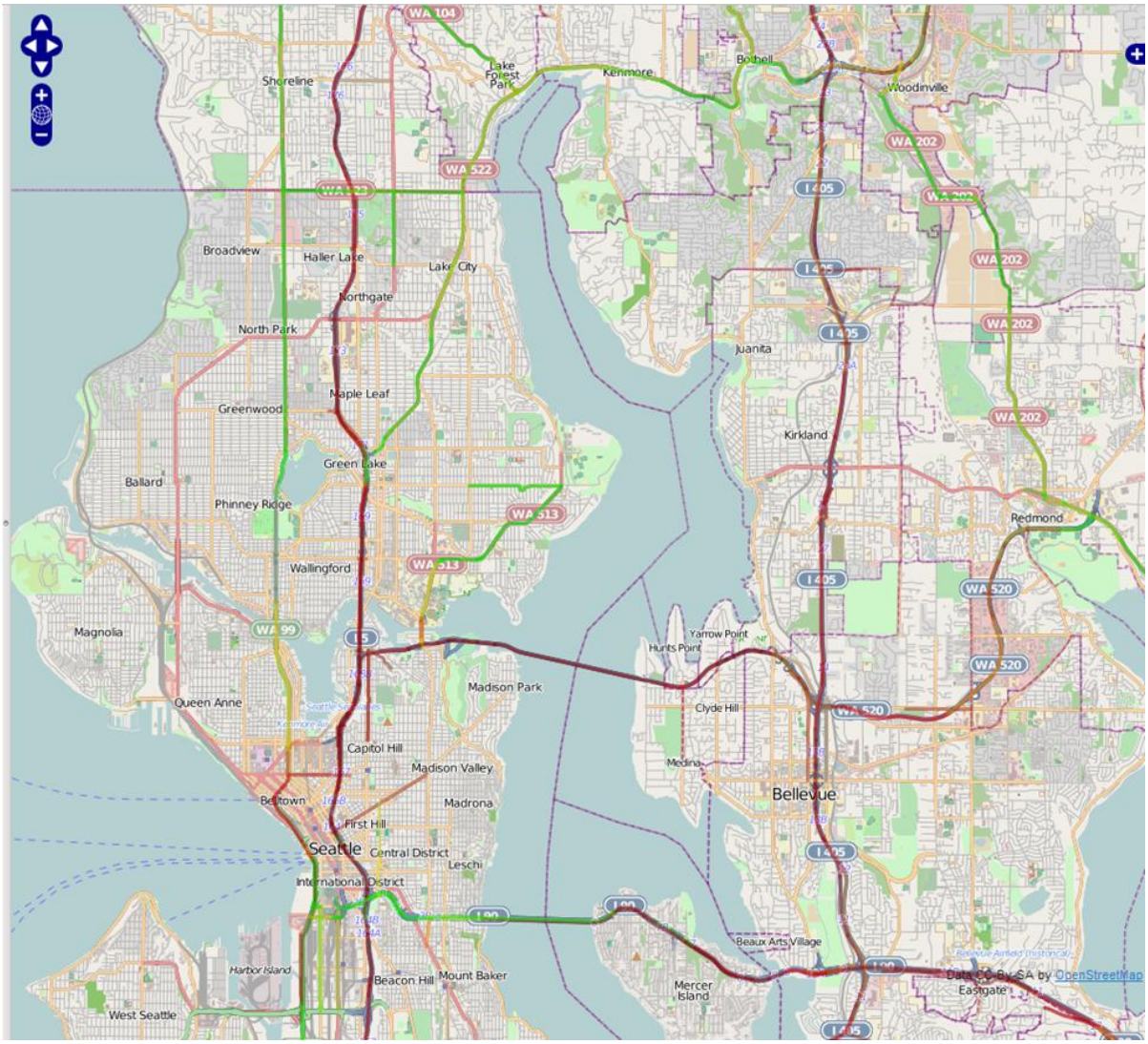


Figure 6-26 Data Visualization: LOS Map

Chapter 7: Conclusion and Future Work

This research presents an eScience transportation platform, namely, DRIVNet 3.0, to model the transportation data. New architecture design is proposed with motivations to facilitate transportation data-driven research. To demonstrate the computational capability of DRIVENet, a pilot study of automating real-time freeway performance measurement is conducted. The study proposed and implemented modeling framework for freeway performance measurement in DRIVENet, enabling users to probe traffic condition timely and avoid time-consuming manual input for analysis. Particularly, it utilizes innovative spatial data fusion technique - pix-based segmentation – to spatially integrate multiple datasets. Applying HCM 2010 methodologies and multi-regime model, the system leverages network-wide datasets to give accurate and real-time evaluation of freeway facilities, which great facilitates users understanding of the traffic data and assists with drawing useful inferences and decision making.

DRIVENet research is still ongoing. System is under constant refinement to keep up with the latest technologies. However, the goals of this research has never changed, which aims to efficiently make big transportation data reliable, accessible, interoperable, and understandable. The contributions from this study are thus summarized as follows:

Reliability

A novel architecture is proposed and implemented with the theme of "thin-client" and three basic tiers, including presentation tier, computational/logic tier and data tier. Comparing with the previous version DRIVENet 2.0, the new system is more robust, supporting a variety of services mining datasets. Moreover, due to the "thin-client" design, no requirement except browsers is needed at the client side, ensuring the system compatibility to the maximal extent. In addition,

forking open source projects improves the reliability of DRIVENet by taking advantage of developer communities all over the world.

Accessibility

The web-based interface makes historical and real-time transportation data as well as a variety of functionalities available through Internet: (1) the user interface is designed in a way that users can easily perceive and interpret; (2) a huge amount of transportation data is currently delivered on the website for use by the transportation researchers, practitioners, and the general public; (3) The system supports various ways for users to explore and exploit data, including downloading, interacting with map, statistically analyzing, visualizing, etc.

Interoperability

The richness of the data provides the resources to probe interdisciplinary and the computational tier enables in-depth data analysis. On the one hand, multidisciplinary data inspires new approaches to reveal insights from data. On the other hand, the flexibility and customizable capability makes DRIVENet an effective tool to develop methods and automate algorithms stand alone, which is fully demonstrated in the case study. The integration of the tool R further equips the system with the latest statistical models from academia.

Understandability

A variety of measurement analysis built in DRIVENet allows users to understand the performance of facilities easily. For instance, the real-time freeway performance measurement introduced in Chapters 5 and 6 qualifies the freeway quality of service into LOS ranging from **A**

through **F**. In addition, tubular and visual analytical tools, as one of key components in DRIVENet, deliver the information in a straightforward and effective fashion.

Future work involves using DRIVENet to provide solutions to other practical problems, such as safety performance assessment, active traffic management decisions, and HOT lane strategy evaluation and optimization, travel time reliability and delay quantification, and congestion analysis. Meanwhile, the DRIVENet team at STARLab is actively seeking collaborations with researchers in other disciplines, which aligns well with the theme that "eScience is about global collaboration in key areas of science and the next generation of infrastructure that will enable it" as Dr. John Taylor stated.

Bibliography

Eaton, C., D. DeRoos, T. Deutsch, G. Lapis, and P. Zikopoulos. *Understanding Big Data*. http://www-01.ibm.com/software/data/bigdata/. Accessed Jun. 9, 2013.

Marcella A. Jr. and Greenfield R. S. *Cyber Forensics: A Field Manual for Collecting, Examining, and Preserving Evidence of Computer Crimes*. CRC Press, 2002.

Gantz J. and D. Reinsel. *Extracting Value from Chaos*. IDC iView, pages 1–12, 2011. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. Accessed Jun. 9, 2013.

Hey T., and A. E. Trefethen. The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems*, Volume 18, Issue 8, October 2002, pp. 1017-1031.

Hey T., and A.E. Trefethen. Cyberinfrastructure for e-Sciecne. *Science*, 308, 6 May 2005, pp. 817-821.

Gianchandani E. *Obama Administration Unveils $200M Big Data R&D Initiative*. http://www.cccblog.org/2012/03/29/obama-administration-unveils-200m-big-data-rd-initiative/. Accessed Jun. 9, 2013.

Ma X., Y. J. Wu, and Y. Wang. DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis. *Transportation Research Record: Journal of the Transportation Research Board.* Vol.2215, pp.37-49, 2011.

Google Inc, *Google Maps API Licensing*, Nov 2012. https://developers.google.com/maps/licensing. Accessed Jun. 9, 2013.

OpenStreetMap. http://www.openstreetmap.org/. Accessed Jun. 9, 2013.

OpenLayers. http://openlayers.org/. Accessed Jun. 9, 2013.

Chen C., K.Petty, A. Skabardonis, P. Varaiya, and Z. Jia. Freeway Performance Measurement System: mining loop detector data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol.1748, pp. 96–102, 2001.

Chao C., J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya. Detecting Errors and Imputing Missing Data for Single-loop Surveillance Systems. *Transportation Research Record: Journal of the Transportation Research Board*, Vol.1855, pp.160–167, 2003.

Tsekeris T., and A. Skabardonis. On-line Performance Measurement Models for Urban Arterial Networks. In *Transportation Research Board 83rd Annual Meeting Compendium of Papers* CD-ROM, 2004.

Petty K., J. Kwon, and A Skabardonis. *A-PeMS: An Arterial Performance Measurement System*. In 2006 Annual Meeting Workshop. Washington, DC, 2005.

CATT Lab. *RITIS System*, 2012. http://www.cattlab.umd.edu/?portfolio=ritis. Accessed Jun. 9, 2013.

Tufte K. A., R. L. Bertini, J. Chee, R. J. Fern′andez-Moctezuma, S. Periasamy, S. Sarkar, P. Singh, J. Whiteneck, S. Matthews, N. Freeman, and S. Ahn. Portal 2.0: Towards a Next Generation Archived Data User Service. In Preprint CD-ROM for *the 89th Annual Meeting of Transportation Research Board*, Washington, DC, 2010.

Xie G., and B. Hoeft. Freeway and Arterial System of Transportation Dashboard. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2271, pp. 45–56, 2012.

Rouphail N., and B Schroeder. *User's Guide to FREEVAL 2010*. http://sites.poli.usp.br/d/ptr2377/HCM2010-FREEVAL%20User%20Guide%20Final_02-27-2011.pdf. Accessed Jun. 9, 2013.

Highway Capacity Manual 2010, Volumes 1 - 4. Transportation Research Board, 2010.

Greenshields B. D., JR Bibbins, WS Channing, and HH Miller. A Study of Traffic Capacity. In *Highway research board proceedings*, 1935.

Pipes L. A.. Car-Following Models and the Fundamental Diagram of Road Traffic *Transportation Research*, Vol. 1, pp. 21–29, 1967.

Drew, D. R. *Traffic Flow Theory and Control*. No. 467, 1968.

Edie, L. C. "Car-following and Steady-state Theory for Noncongested Traffic." *Operations Research 9.1*, pp. 66-76, 1961

Underwood R. T.. Speed, Volume and Density Relationships. *Quality and Theory of Traffic Flow Yale Bureau of Highway Traffic*, New Haven, Connecticut, pp. 141–188, 1961.

Greenberg H.. An Analysis of Traffic Flow. *Operations Research*, Vol. 7, pp. 79–85, 1959.

Drake J., J. Schofer, and A. May. A Statistical Analysis of Speed-density Hypotheses. *Traffic Flow and Transportation*, 1965.

Sun L., and J. Zhou. Development of Multiregime Speed-density Relationships by Cluster Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1934, pp. 64–71, 2005.

MacQueen, J.. Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297, 1967.

Steinhaus, H.. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1:801–804, 1956.

Yu R., Y. Lao, X. Ma, and Y. Wang. Short-term Traffic Flow Forecasting for Improved Estimates of Freeway Incident Induced Delays. *In 90th Annual Meeting of the Transportation Research Board*, Washington, DC, 2011.

Malinovskiy Y., N. Saunier, and Y. Wang. Pedestrian Travel Analysis Using Static Bluetooth Sensors. *In Transportation Research Board 91st Annual Meeting*, No. 12-3270, 2012.

Ma X., E. D McCormack, and Y. Wang. Processing Commercial Global Positioning System Data to Develop a Web-based Truck Performance Measures Program. Transportation Research Record: *Journal of the Transportation Research Board*, Vol.2246, pp. 92–100, 2011.

Lewandowski, S. M. Frameworks for Component-based Client/Server Computing. *ACM Computing Surveys* (CSUR) 30.1, pp.3-27, 1998.

Vaadin. https://vaadin.com/book/vaadin7/-/page/intro.html. Accessed Jun. 9, 2013.

Wang, Y., et al. *Development of a Statewide Online System for Traffic Data Quality Control and Sharing*. No. TNW2009-12. 2009.

Wang Y., M. Hallenbeck, P. Cheevarunothai, and Transportation Northwest. *Quantifying incident-induced travel delays on freeways using traffic sensor data*. Technical report, Transportation Northwest, University of Washington, 2008.

Haklay, M., and P. Weber. OpenStreetMap: User-generated Street Maps. *Pervasive Computing, IEEE 7.4*, pp.12-18, 2008.

Goodchild, M. F. Citizens as sensors: the world of volunteered geography. *GeoJournal 69.4*, pp. 211-221, 2007.

Goodchild, M. F. Commentary: whither VGI?. *GeoJournal* 72.3 pp. 239-244, 2008.

Wood H.. *1 million OpenStreetMappers*, January 2013. http://blog.openstreetmap.org/2013/01/06/1-million-openstreetmappers/. Accessed Jun. 9, 2013.

Microsoft, Bing Blogs. *Bing Engages Open Maps Community*, November 2010. http://www.bing.com/blogs/site_blogs/b/maps/archive/2010/11/23/bing-engages-open-maps-community.aspx. Accessed Jun. 9, 2013.

OpenStreetMap. *Copyright and License*. http://www.openstreetmap.org/copyright. Accessed Jun. 9, 2013.

Zielstra D., and A. Zipf. A comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *13th AGILE International Conference on Geographic Information Science*, volume 2010, 2010.

Zielstra, D., and H. H. Hochmair. Digital Street Data: Free versus Proprietary. GIM Int 25 pp. 29-33, 2011.

Haklay M.. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and planning. B, Planning & design*, 37(4):682, 2010.

Haklay M, A. Singleton, and C. Parker. Web Mapping 2.0: The Neogeography of the Geoweb. *Geography Compass*, 2(6):2011–2039, 2008.

Obe R., and L. Hsu. *PostGIS in Action*. Manning Publications Co., 2011.

PostgreSQL. http://www.postgresql.org/about/. Accessed Jun. 9, 2013.

pgRouting. http://pgrouting.org/. Accessed Jun. 9, 2013.

R. http://www.r-project.org/. Accessed Jun. 9, 2013.

Rserve. http://www.rforge.net/Rserve/. Accessed Jun. 9, 2013.

Kittelson, W. K. Historical overview of the committee on highway capacity and quality of service. *Transportation Research Board-TRB, National Research Council-Transportation Research Circular E-C018: 4th International Symposium on Highway Capacity*, Maui, Hawaii June. 2000.

Wikipedia. Image Resolution. http://en.wikipedia.org/wiki/Image_resolution. Accessed Jun 9, 2013.

WSDOT. *WSDOT's Linear Referencing System*. http://www.wsdot.wa.gov/mapsdata/tools/traffictrends/tptappendicesforwsdotlrs.pdf. Accessed Jun. 9, 2013.

Roadway Datamart for GIS. http://www.wsdot.wa.gov/mapsdata/geodatacatalog/Maps/noscale/DOT_TDO/RoadwayDatamart/RoadwayDatamartIDX.htm. Accessed Jun. 9, 2013.

Zegeer, J. D., M. Vandehey, M. Blogg, K. Nguyen, and M. Ereti. NCHRP Report 599: Default Values for Highway Capacity and Level of Service Analyses. *Transportation Research Board of the National Academies*, Washington, DC, 2008.

Turner S.M.. *Guidelines for Developing ITS Data Archiving Systems*. Technical report, 2001.

Jacobson L. N., N. L Nihan, and J. D Bender. Detecting Erroneous Loop Detector Data *In a Freeway Traffic Management system*. Number 1287. 1990.

Chen C., J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya. Detecting errors and imputing missing data for single-loop surveillance systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1855(1):160–167, 2003.

Ishak S.. Fuzzy-clustering approach to quantify uncertainties of freeway detector observations. *Transportation Research Record: Journal of the Transportation Research Board*, 1856(1):6–15, 2003.

Kwon J., C. Chen, and P. Varaiya. Statistical methods for detecting spatial configuration errors in traffic surveillance sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 1870(1):124–132, 2004.