Research Report
Agreement T1461, Task 05
DRIVE Net 2

DIGITAL ROADWAY INTERACTIVE VISUALIZATION AND EVALUATION NETWORK APPLICATIONS TO WSDOT OPERATIONAL DATA USAGE

by

Yinhai Wang
Professor

Weibin Zhang
Research Associate

Kris Henrickson
Graduate Research Assistant

Ruimin Ke
Graduate Research Assistant

Zhiyong Cui
Graduate Research Assistant

Smart Transportation Applications and Research Laboratory (STAR Lab)
Department of Civil and Environmental Engineering
University of Washington
Seattle, Washington 98195-2700

Washington State Department of Transportation Technical Monitor
Bill Legg, State ITS Operations Engineer

Prepared for

The State of Washington
Department of Transportation
Roger Millar, Acting Secretary

December 2016

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
WA-RD 854.1		

4. TITLE AND SUBTITLE	5. REPORT DATE
DIGITAL ROADWAY INTERACTIVE VISUALIZATION AND EVALUATION NETWORK APPLICATIONS TO WSDOT OPERATIONAL DATA USAGE	December 2016
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Yinhai Wang, Weibin Zhang, Kristian Henrickson, Ruimin Ke, Zhiyong Cui	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NO.
Smart Transportation Applications and Research Laboratory Box 352700, 101 More Hall University of Washington Seattle, WA 98195-2700	11. CONTRACT OR GRANT NO. Agreement T1461, Task 05

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Washington State Department of Transportation Transportation Building, MS 47372 Olympia, Washington 98504-7372 14 Doug Brodin, Project Manager, 360-705-7972	Research Report 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This study was conducted in cooperation with the University of Washington.

16. ABSTRACT

DRIVE Net is a region-wide, Web-based transportation decision support system that adopts digital roadway maps as the base, and provides data layers for integrating and analyzing a variety of data sources (e.g., traffic sensors, incident records). Moreover, DRIVE Net offers a platform for streamlining transportation analysis and decision making, and it serves as a practical tool for visualizing historical observations spatially and temporally. In its current implementation, DRIVE Net demonstrates the potential to be used as a standard tool for incorporating multiple data sets from different fields and as a platform for real-time decision making. In comparison with the previous version, the new DRIVE Net system is now able to handle more complex computational tasks, perform large-scale spatial processing, and support data sharing services to provide a stable and interoperable platform to process, analyze, visualize, and share transportation data.

DRIVE Net's capabilities include generating statistics for WSDOT's Gray Notebook (GNB), including travel times, throughput productivity, and traffic delay calculations for both general purpose and HOV lanes, each of which are important performance indicators in the WSDOT congestion report. The DRIVE Net system includes robust loop detector data processing and quality control methods to address the data quality issues impacting loop detectors throughout the state. The capabilities of the DRIVE Net system have been expanded to include safety modeling, hotspot identification, and incident induced delay estimation. Specifically, the Safety Performance module includes functions that can be used to obtain traffic incident frequency, apply predictive models to estimate the safety performance of road segments, and visualize and compare observed incident counts and different predictive models. Additionally, a module providing multi-modal data analysis and visualization capabilities was developed as a pilot experiment for integration of heterogeneous data. This module includes pedestrian and bicycle, public transit, park and ride, Car2Go, and ferry data downloading and visualization. DRIVE Net now offers role-based access control, such that access privileges to different functions and data resources can be assigned on a group or individual basis.

The new system is able to support more complex analytics and decision support features on a large-scale transportation network, and is expected to be of great practical use for both traffic engineers and researchers. With a modular structure and mature data integration and management framework, DRIVE Net can be expanded in the future to include a variety of additional data resources and analytical capabilities.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
Freeway performance measurement, WSDOT Gray Notebook, geospatial data fusion, data Quality Control, automatic pedestrian data collection	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22616

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
None	None		

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the data presented herein. This document is disseminated through the Washington State Department of Transportation. The contents do not necessarily reflect the views or policies of Washington State Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

Table of Contents

Table of Figures

Table of Tables

Executive Summary

In recent years, the increasing quantity and variety of data available for decision support present a wealth of opportunity as well as a number of new challenges, in both the public and private sectors. Vast quantities of data are available through increasingly affordable and accessible data acquisition and communications technologies, including sensors, cameras, mobile location services, and others, which, when combined with emerging computing and analysis methodologies, can lead to more informed and proactive management and decision making. As a result, big data concepts and methodologies are steadily moving into the mainstream in a variety of science and engineering fields.

In past decades, transportation research has been largely driven by mathematical equations and has relied on relatively scarce data. With the increasing quantity and variety of data being collected from intelligent transportation systems and other sensors, data-driven or data-based research is rapidly growing more possible. Nevertheless, today there are few established systems for applying big data tools to transportation analysis and decision support. Most current online data analysis and visualization systems are designed to handle one type of data, such as from freeway or arterial sensors. Therefore, though the scope and ubiquity of transportation data are increasing, making these data accessible and useable for transportation analysis is a difficult challenge.

Most commercial systems are oriented toward a specific transportation problem or analysis procedure, and approach the problem in their own, often ad hoc, way. A mature framework for effectively utilizing data and computing resources, such that these data will serve the needs of users, has become a pressing need in the field of transportation. The challenges associated with developing this type of framework primarily stem from the need for standardized and efficient data integration and quality control methods, computational modules for applying these data to transportation analysis, and a unified data schema for heterogeneous data.

The goal of this study was to remove the barriers in the current data sets that the Washington State Department of Transportation (WSDOT) archives, automate its time-consuming data quality control process, and achieve the integration and visualization of information necessary to support decision making. The resulting data fusion techniques and

database design details are both summarized in this report and delivered in a functioning online system named WSDOT Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net). This WSDOT DRIVE Net system is capable of ingesting, archiving, and quality checking traffic sensor data from all WSDOT regions. It has also been built to incorporate data from a range of WSDOT and third party sources, including linking travel times from INRIX and the National Performance Management Research Data Set (NPMRDS), the Washington Incident Tracking System (WITS), and roadside weather sensors. Unlike other transportation data by archiving systems, DRIVE Net is also capable of processing and managing massive amounts of spatial data using open-sourced spatial database tools. This will significantly alleviate the computational and financial burden of using commercial geographic information system (GIS) software packages and will provide maximum flexibility to end users. Traditional transportation and spatial data are integrated into a robust GIS-T model for large-scale modeling and network-level performance estimation following the principles of eScience.

DRIVE Net is a region-wide, web-based transportation decision support system that adopts digital roadway maps as the base and provides data layers for integrating a variety of data sources. Moreover, DRIVE Net offers a platform for streamlining transportation analysis and decision making, and it serves as a practical tool for visualizing historical observations both spatially and temporally. In its current implementation, DRIVE Net demonstrates the potential to be used as a standard tool for incorporating more data sets from different fields and as a platform for real-time decision making. In comparison with the previous version, the new DRIVE Net system is now able to handle more complex computational tasks, perform large-scale spatial processing, and support data sharing services to provide a more stable yet interoperable platform to process, analyze, visualize, and share transportation data.

By leveraging the power of eScience, DRIVE Net can be used to quickly and automatically generate statistics for WSDOT's Gray Notebook (GNB), including travel time analysis, throughput productivity evaluation, travel time delay calculations, and HOV analysis, each of which is important performance indicator in the WSDOT congestion report. The capabilities of the DRIVE Net system have been expanded to include traffic crash modeling and hotspot identification. Specifically, the Safety Performance module includes functions that can

be used to obtain traffic incident frequency, apply predictive models to estimate the safety performance of different road segments, and visualize and compare observed incident counts and different predictive models. Additionally, a module providing multi-modal data analysis and visualization capabilities, developed as a pilot experiment, includes pedestrian and bicycle, public transit, park and ride, Car2Go, and ferry data. This is a meaningful attempt to integrate different data sources within the single platform. At present, the module is able to provide multi-source data download and visualization, and more features are expected in the near future.

Underlying the analysis modules, a crucial component of the DRIVE Net system is the loop detector data processing and quality control methodology. WSDOT manages a great number of loop detectors throughout the state, and significant amounts of transportation analysis and reporting relies on the quality of loop detector observations. Although WSDOT provides a preliminary data quality assurance procedure to flag erroneous loop data, this procedure is still unable to capture other possible errors, such as loop detector sensitivity issues. Because of the environmental changes around loop detectors over time, the actual detection zone of these loops may increase or decrease, and these changes will consequently affect the accuracy of speed calculations. It is of critical importance to detect and correct possible loop errors before freeway performance measurement is conducted. This study developed and implemented a robust loop data quality control mechanism in DRIVE Net.

In summary, this work describes the development of an eScience transportation platform and provides an interoperable, data-driven online alternative to WSDOT's existing data systems. Significant contributions are listed below:

(1) The DRIVE Net system capability has been significantly enhanced with the development of multiple new modules, the ability to import different data sources, and database design optimization.
(2) This study implemented an efficient and effective GIS-T model to integrate and visualize immense amounts of transportation data from various sources into a single platform.
(3) By incorporating more heterogeneous data sets, including not only loop data, GPS speed and travel time data, weather data, and WITS data, but also pedestrian and

bicycle, public transit, park and ride, Car2Go, and ferry, this study has improved agencies' ability to achieve more comprehensive and robust transportation analysis.

(4) Production for the WSDOT Gray Notebook has been included as a key component in the DRIVE Net system. Raw loop data are automatically processed in a series of rigorous data quality control procedures and are available in a number of different analysis modules, such as travel time analysis, throughput productivity evaluation, and travel time delay estimation for both general purpose and high occupancy vehicles.

(5) DRIVE Net now includes enhanced travel time reliability estimation and visualization functions. Specifically, users can estimate travel time reliability measures for both predefined and custom corridors and can visualize maximum travel distances for fixed travel times at various levels of reliability.

(6) The safety analysis module includes predictive modeling, hotspot identification, and various visualization tools for assessing the safety performance of the highway system. Furthermore, users can view specific incidents, calculate the associated traffic delay, and visualize and compare different locations and time periods.

(7) A multi-modal analysis module has been developed that includes data download and visualization of data from multiple sources.

(8) DRIVE Net now features enhanced data sharing capabilities with role-based access control. As new data sources are added to DRIVE Net, WSDOT will be able to assign access to different functions and data sets on an individual or group basis.

With a modular structure and mature data integration and management framework, DRIVE Net can be expanded in the future to include a variety of additional data resources and analytical capabilities. Future developments to the scope and power of DRIVE Net will include incorporating data from new and emerging technologies, enhanced multi-dimensional data integration, and richer travel analysis functions.

Chapter 1 Introduction

1.1 Problem Statement

The Washington State Department of Transportation (WSDOT) has a significant need for a tool that has the data storage and analysis capability to allow fast analysis of data from multiple sources in support of WSDOT's current and future *Moving Washington* strategies. This includes support of project planning, scoping, design, construction, performance analysis, reporting, system maintenance, and active traffic management activities. Modern technology is creating a significant increase in the amount as well as the types of data available to describe the condition, use, and performance of the State's transportation system. For instance, traffic detectors have been widely deployed over the state highway network in Washington. Third party data, such as Inrix GPS data, and HERE data, complement WSDOT traffic sensor network data, particularly for rural areas where traffic detectors are sporadic. The combined WSDOT and third party data are huge in volume and highly valuable for system operations, monitoring, and analysis.

While many new data sources are being captured, these data sets are not being used to WSDOT's full benefit because they cannot be easily combined with each other or be integrated into WSDOT's existing data systems. The reason is that most of the current traffic data archive systems were designed mainly for data storage and off-line analysis (Ma et al., 2011). They lack capabilities that could integrate third party data sets for analysis and are not capable of offering the functions needed for real-time performance monitoring, quick operational decision support, and system-wide analysis. Additionally, the recently deployed Intelligent Transportation Systems (ITS) technologies in Washington State, such as active traffic management (ATM), and high occupancy toll (HOT) lane operations, require real-time performance monitoring and reliable performance reporting to maximize their utilities. System-wide performance reports are required for WSDOT's accountability reporting and management functions. Currently, performance monitoring and analysis are both labor intensive and time consuming because of a lack of modern tools, thus making these reports slow and costly to deliver and limiting WSDOT's ability to be responsive to legislative and agency requests for information.

To address these problems, the latest version of the Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) system, developed by the University of Washington, has been expanded from the previous version to include additional data sources and new, desired analytical functions. The expended DRIVE Net is a great platform for various kinds of analysis.

1.2 General Background

This research was based on the DRIVE Net Phase I research at the University of Washington. It established an eScience transportation platform capable of integrating various data sets on a regional map basis. Since roadway geometric data, traffic sensor data, incident data, and more are tied together on this platform, it offers a variety of functions for large-scale system operations, monitoring, and analysis, such as travel time analysis, safety analysis, and emissions calculation. In this Phase II project, the coverage breadth and analysis depth of DRIVE Net were increased. Specifically, DRIVE Net now supports the following functions:

- Operational strategy evaluation and optimization

  Impacts of a new technology (or construction) likely occur beyond the specific deployment site. For example, if an ATM corridor demonstrated improvements in travel time reliability and safety, more travelers might be attracted to the corridor, offsetting the benefit observable at its original demand level but creating benefits on the parallel facilities that were losing traffic to the ATM-equipped roadway. Similarly, major reconstruction efforts (and the mitigation associated with those efforts) have system-wide implications that need to be tracked and reported. Such dynamic features within traffic networks require analysis over a sufficiently large scale for a reliable assessment of before-and-after conditions. Systematic analysis is also needed to optimize the configuration of operational strategies and to estimate how much throughput can be recovered to enhance WSDOT's ability to evaluate proposed ATM/HOT lanes and other operational efficiency initiatives and to advocate for them. This operational strategy evaluation and optimization function can be implemented by defining relevant performance measures and computing those measures efficiently over the user-defined

network. This function can be used to quantify the impacts of operational strategies and identify directions for improvement.

- Statewide/regional/corridor-level systems analysis

  The ability to analyze the available data to measure performance metrics across the state, rather than limit analysis to the Puget Sound area, is of utmost importance, as often the results of these analyses become elected officials' speaking points on WSDOT's system performance. This analysis capability takes into account the best available data from different sources. As emphasized by Moving Ahead for Progress in the 21st Century (MAP-21), freeway congestion/system performance measures are critical for transportation system operations. The key performance indicators include percentage of system delay occurring within specified boundaries, amount of congestion occurring at specific locations, and total vehicle hours of delay.

- Identification of existing geometric design or roadway control problems

  For example, WSDOT staff periodically review the appropriateness of speed limits set on state routes throughout the state highway network. Such a review often requires long-term performance observations and analyses, which have been challenging to conduct with the previous data archive system. The new system is able to provide decision support that takes historical traffic and collision data into account to provide guidance on applicable solutions.

- Informed decision making

  Existing data must be better utilized to understand network performance and the impacts that different factors have on that performance. For example, where do bottlenecks form and how do they evolve? What are the likely causes of the recurrent bottlenecks? How do those bottlenecks change as new operational controls are implemented? Which segments are subject to long incident delays and demand more incident response resources? How do drivers react to tolls, both on the tolled roads and on alternative routes?

- Safety performance assessment

Most traffic crash modeling and safety performance analyses are based on spreadsheet data in which observed traffic crashes are categorized in geometrically uniform roadway segments.

Such analyses cannot capture their impacts of dynamic factors, such as flow variation and road surface condition over time, although those factors are often critical for understanding the causes of crashes and assessing the effectiveness of safety countermeasures. Integrating mobility data into system-wide safety analyses will be helpful for identifying accident-prone locations and their contributing factors.

1.3 Research Objectives

This project aimed at developing data-driven methods and tools to help WSDOT's decision making and operational practices. Toward this project goal, the following objectives were set:

- Expand the scope of DRIVE Net to cover the entire state highway network covered by existing data.
- Develop data fusion algorithms to extract performance measures from traffic sensor data and private sector spot speed data.
- Develop a function for travel delay quantification and congestion analysis. The quantification of travel delay should be scalable on the basis of times of day and geography (statewide level, region-wide, NHS system to corridor level). This analytical module should have the capability to develop other delay related metrics such as delay per capita, delay per commuter, and delay per peak period traveler. In addition, the analytical module should provide flexibility to use (input) different threshold values to define delay, while the WSDOT standard threshold for delay definition is less than 85 percent of posted speed.
    - WSDOT is also interested in integrating transit ridership, bicycling, and walking data sets where available with the DRIVE Net data sources so that they can be judiciously integrated into performance measurement calculations related to delay avoided by transit and non-motorized transportation when multimodal performance is measured.
    - Within the delay module it is important to be able to compute VMT so that it can be used to compare performance measure trends for deeper analysis purposes.

These performance indicators will be used to carry out the congestion mitigation and air quality improvement programs proposed by MAP-21.

- Improve current crash modeling methods by introducing a non-linear prediction function in the Generalized Nonlinear Models (GNMs) to describe the relationship between injury severity and its contributing factors (including but not limited to weather conditions, ITS operations, and roadway geometrics).
- Develop a multimodal data download and visualization panel, integrating different data sources into the DRIVE Net system, including park and ride, ferry, and bicycle counting data.
- Develop a Safety Performance Index (SPI), based on the expected accident frequency derived from the improved modeling approach and accident associated factors, to reflect safety conditions related to changes with roadway, vehicle, mobility, and environmental factors.
- Enhance visualization capabilities and add customized reporting functions to integrate more relevant WSDOT planning, programming, and operation activities into the DRIVE Net.

Chapter 2 Literature Review

Over the last few years, the volume of available data has exploded. With the fast development of networking, data storage, and data collection by new sensors, big data, representing a new era in data exploration and utilization, is now rapidly expanding into the transportation arena. Big data give us opportunities to better monitor traffic networks and to increase the accuracy of traffic predictions. However, the methods, models, and algorithms that are used today in the transportation domain to mine and explore data may not scale and/or perform well under these new conditions (Vlahogianni et al. 2015). In facing the challenges of large data volumes and rapid data generation, we need to develop novel ways to manage and analyze big data.

Transportation data accessibility and usability for the public and policy makers are important. According to Chen et al. (2015), traffic data types determined by sensor mode can be categorized into three classes: location-based, activity-based and device-based. A sequence of data preprocessing operations is required before the data are ready to use, including data cleaning, data matching, data organization, and data aggregation. Given the numerous application program interfaces (APIs), such Google Map APIs (Google, 2016), Uber APIs (Uber 2016), OneBusAway APIs (OneBusAway 2016), and local agency APIs that may provide data, the downloaded data are in many different formats and types and not available for further comprehensive analysis. Khazaei et al. (2015) proposed a platform for performing analytics on urban transportation data with a three-layer data management subsystem. Although this platform integrates data from loops, traffic cameras, incident reports, and mobile devices, the analytic data are independent, and it is therefore not easy to extend their analytics to answering questions requiring multi-source data posed by different types of users, such as transportation managers, traffic engineers, planners, researchers, and policy makers. Therefore, data aggregation and quality control are extremely necessary before the data can be interpreted, used, and shared.

Data visualization is an efficient way to discover regular patterns or abnormal features when huge amounts of data are analyzed and used to build models. Specifically, traffic data visualization can facilitate understanding of the behavior of moving objects and discovery of traffic, social, geo-spatial, and even economic patterns. Chen et al. (2015) conducted a survey of

traffic data visualization and introduced the basic concepts of traffic data visualization. They also summarized the existing state-of-the-art methods for depicting the temporal, spatial, numerical, and categorical properties of traffic data, which have been adopted by many outstanding traffic data projects (Kloeckl et al. 2016; Barry et al. 2016) and intelligent transportation systems (ITS) (Pu et al. 2013). With the development of data visualization techniques, transportation practitioners and researchers have made great progress in traffic data visualization. Shekhar et al. (2002) modeled traffic data and established a web-based system for traffic data visualization (CubeView). A visualization tool (MetroViz) developed by Du et al. (2015), which aims to help users explore public transportation data and evaluate the performance of public transportation systems, has much better interactive visualization performance. Comparing previous transportation data visualization tools, MetroViz can provide an overview of data that are interactively selected by users. Many traffic visualization projects and systems have also been developed to focus on specific types of transportation data, such as transit data (Du et al. 2015), trajectory data (Scheepens et al. 2016; Wang et al. 2014), taxi data (Huang et al. 2016) and bike sharing system data (Bargar et al. 2014).

The availability of big data can potentially lead to a revolution in ITS development and related research. ITS is changing from conventional, technology-driven systems into more powerful, multifunctional, data-driven systems (D^2ITS). Zhang et al. (2011) conducted a survey on the development and deployment issues related to D^2ITS and introduced several of its important components, including vision-, multisource-, and learning-driven ITS. By using D^2ITS, researchers and policy makers can augment the existing models and make better decisions. Furthermore, a more attractive advantage of big data is that new products and services will emerge on the basis of well-developed, efficient, and user-friendly ITS platforms. Microsoft Research has dedicated extensive efforts to analyzing large-scale data, carrying out influential research, and introducing them into practice, such as diagnoses of urban noise based on big data (Zheng et al. 2014), a passenger-cabbie recommender system (Yuan et al. 2011), and a large-scale dynamic taxi ridesharing service (Ma et al. 2013). GE has made a major investment in new service models for its industrial products using big data analytics. In the Las Vegas and Nevada metropolitan areas, the Freeway and Arterial Systems of Transportation dashboard (FAST), a

web-based system, controls and monitors traffic by integrating traffic data (lane occupancy, volume, and speed), incident data and weather data (Xie and Hoeft, 2012). The Arterial Performance Measurement System (APeMS) has been implemented to estimate intersection travel time, control delay, and progression quality on arterials every 5 minutes on the on the basis of mid-block loop detector data (Tsekeris et al., 2004; Petty et al., 2005). Hitachi is developing city management platform solutions that support efficient city operations by analyzing big data collected from the public and traffic flows (Morioka et al. 2015). The Regional Integrated Transportation Information System (RITIS) in Maryland can automatically fuse, translate, and standardize data obtained from multiple agencies in the National Capital Region in order to provide an enhanced overall view of the region's transportation network (Pack et al. 2008). However, despite the numerous data sharing and visualization techniques and platforms that have been developed, there is still room for considerable improvements and opportunities to create a more efficient and widely-used e-Science transportation platform for data sharing, visualization, and analysis.

Chapter 3 Study Data

DRIVE Net builds upon existing databases controlled by the STAR Lab. A variety of data sources are digested by and archived into the STAR Lab server from WSDOT and third party data providers through different data acquisition methods. There are four ways to use the data archive service:

1. Direct upload

Users can upload data into the database through the DRIVE Net website. This method is suitable for receiving data from those who do not maintain online databases. Typical study data sets include INRIX data and weather station data.

2. Periodic download via Web services

A scheduled fetch job is run to download data at predefined intervals via File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP), Simple Object Access Protocol (SOAP), or Representational State Transfer Principles (RESTful) interfaces. This method is currently used for the acquisition WSDOT freeway loop data.

3. Active data acquisition

For those agencies with specialized needs or that do not allow public access, the research team will construct a satellite server—a form of "information appliance"—that bundles hardware, software, and data processing services into a single provisionable platform. These satellite servers elegantly solve several problems related to bootstrapping a data sharing network. First, system administrators rarely create holes in their firewalls for connections with remote machines. The appliance, however, can be deployed inside the agency's firewall and still connect to remote servers by using port 80 or port 22, which are usually unrestricted. Second, specialized software for establishing a Web service, in order to use the periodic download method, is difficult to install and configure. Even if a comprehensive software suite is written, the cost of providing technical support to users would be prohibitive. However, installing the software on behalf of a customer on computers over which the STAR Lab has complete control is far more

straightforward. Finally, the appliance grants access to STAR Lab researchers and technicians as well as participant agency staff. This allows multi-agency shared access, which can simplify troubleshooting and upgrade deployment. This method is currently used to retrieve the roadway geometric data and WITS data from WSDOT.

4. Direct data archiving

Data are generated from data collection devices and enter into the data warehouse by several communication protocols, such as General Packet Radio Service (GPRS) and Global System for Mobile Communications (GSM). Mobile sensor data are transmitted into DRIVE Net with this method.

Detailed information about each data source is described in the following sections.

3.1 Freeway Loop Data

Inductive loop detectors are widely used to monitor freeway performance in the United States because of their reliability and durability (Klein et al, 2006). An inductive loop detector is a conductive coil embedded in the pavement, and it detects a moving vehicle passing over it with electromagnetics. The signal is then transmitted to a roadside cabinet, which stores the vehicle presence information and also sends the signal to the traffic management center via cable. Volume and occupancy are two key indicators that traffic detectors can collect during a fixed time interval (20 seconds or 5 minutes). WSDOT maintains and manages loop detectors in both Washington state highway and Interstate freeways. Washington divides the state into six regions: Northwest, North Central, Eastern, South Central, Southwest, and Olympic. For instance, approximately 4200 single or dual loop detectors are installed in the Northwest Region, and they aim to monitor traffic conditions around the Seattle metropolitan area.

WSDOT stores both 20-second and 5-minute loop detector data by using an online FTP website for downloading. The 5-minute loop detector data are aggregated from 20-second loop data for long-term analysis and archiving. A computer program written in Microsoft Visual C# was developed to periodically retrieve loop data from the posted FTP website, and the downloaded data are automatically imported into Microsoft SQL server databases for further processing.

Single loop detectors can detect only whether a vehicle is present or absent. When several vehicles pass over a single loop detector during a certain time interval, the detector is able to count the number of vehicles and the percentage of time when the detector is occupied. Unlike single loop detectors, a dual loop detector is composed of two single loop detectors, which are placed a short distance apart. By measuring the arrival time difference between the two loops, the roadside traffic controller can calculate each vehicle's speed. The vehicle's length can be also estimated by using the calculated vehicle speed and the on-time measurement from either the front loop or the rear loop.

For both 20-second and 5-minute data aggregation intervals, three types of loop data are collected. The key information is listed in Table 3-1 and Table 3-2.

Table 3-1 20-Second freeway loop data description

Table: SingleLoopData and StationData (Single Loop)

Columns	Data Type	Value Description
LOOPID	smallint	Unique ID number assigned in order of addition to LoopsInfo table
STAMP	datetime	24-hour time in integer format as YYYYMMDD hh:mm:ss (in 20-second increments)
DATA	tinyint	Indicate whether a record is present or not
FLAG	tinyint	Validity flag (0-7): 0=good data; otherwise, bad data
VOLUME	tinyint	Integer volume observed during this 20-second interval
SCAN	smallint	Number of scans when a loop is occupied during each period (60 scans per second multiplied by 20 seconds per period equals 1200 scans)

Table: TrapData (Dual Loop)

Columns	Data Type	Value Description
SPEED	smallint	Average speed for each 20-second interval (e.g., 563 means 56.3 mile per hour)

LENGTH	smallint	Average estimated vehicle length for each 20-second interval (e.g., 228 means 22.8 feet)

WSDOT primarily uses the 5-minute aggregation level loop data for freeway performance measures (Wang et al., 2008). The key information for 5-minute loop data is shown in Table 3-2.

LoopID is the unique ID that matches each cabinet with loop data. Several loops could connect to each cabinet. For each cabinet, these loop data are aggregated as a loop group, namely a loop station, for which the volume is the sum of total volumes for the associated loops, and the occupancy (or scan) is the average of total occupancies (scans) for the associated loops. In addition, to facilitate locating and categorizing each loop, each loop is assigned to a cabinet with spatial information (e.g., milepost). The key information is listed in Table 3-3.

Table 3-2 5-Minute freeway loop data description

Table: STD_5Min and STN_5Min (Single Loop)		
Columns	**Data Type**	**Value Description**
LOOPID	smallint	Unique ID number assigned in order of addition to LoopsInfo table
STAMP	datetime	24-hour time in integer format as YYYYMMDD hh:mm:ss (increased by 5 minutes)
FLAG	tinyint	Good/bad data flag with 1 = good and 0 = bad (simple diagnostics supplied by WSDOT)
VOLUME	tinyint	Integer volume observed during each 5-minute interval
OCCUPANCY	smallint	Percentage of occupancy expressed in tenths to obtain integer values (6.5% = 65)
PERIODS	smallint	The number of 20-second readings incorporated into this 5-minute record (15 is ideal, less than 15 almost always indicates that volume data are unusable unless adjusted to account for missing intervals).

Table: TRAP_5Min (Dual Loop)

Columns	Data Type	Value Description
SPEED	smallint	Average speed for each 5-minute interval (e.g., 563 means 56.3 mile per hour)
LENGTH	smallint	Average estimated vehicle length for each 5-minute interval (e.g., 228 means 22.8 feet)

Table 3-3 Cabinet Data Description

Columns	Data Type	Value Description
CabName	varchar	Unique ID for each cabinet
UnitType	varchar	Type for each loop (i.e. main, station, speed and trap)
ID	smallint	Unique ID number assigned in order of matching the loop data table
Route	varchar	The state route ID (e.g. 005=Interstate 5)
direction	varchar	Direction of each state route
isHOV	tinyint	Bit indication whether loop detector is on an HOV lane (1=HOV, 0=not HOV)
isMetered	tinyint	Bit indication whether loop detector is on a metered ramp (1=metered, 0=not metered)

Although WSDOT provides a preliminary data quality assurance procedure to flag erroneous loop data, this procedure is still unable to capture other possible errors, such as loop detector sensitivity issues (Corey et al., 2011). Because of the environmental changes around loop detectors over time, the actual detection zone of these loops may increase or decrease, and these changes will consequently affect the accuracy of speed calculations. Zhang et al. (2003) stated that approximately 80 percent of WSDOT dual-loops suffer from severe sensitivity problems. It is of critical importance to detect and correct possible loop errors before freeway

performance is measured. A detailed loop data quality control mechanism will be discussed later in this report.

3.2 INRIX Data

As a leading traffic data provider, INRIX combines multiple data sources, including GPS-equipped devices and cell phones. INRIX tracks more than 30 million probe vehicles and more than 400 additional data sources (INRIX, 2012). To aggregate and fuse heterogeneous transportation data, INRIX developed a series of statistical models to compute real-time traffic information such as speed and travel time on the basis of measurements from GPS devices, cellular networks, and loop detectors. The resulting speed data were aggregated into 5-minute intervals for 2008, 2009, and 2010 and into 1-minute intervals for 2011 and 2012. WSDOT purchases the data, and they are further archived into the database in the STAR Lab. INRIX data cover almost the entire roadway network in Washington, including freeways, highways, and most arterials and side streets. The key information for INRIX data is presented in Table 3-4.

Table 3-4 INRIX Data Description

Columns	Data Type	Value Description
DateTimeStamp	datetime	24-hour time in integer format as YYYYMMDD hh:mm:ss
SegmentID	varchar	Unique ID for each segment-Traffic Message Channel (TMC) code
Reading	smallint	Average speed for each segment

INRIX has adopted the Traffic Message Channel (TMC), a common industry convention developed by leading map vendors, as its base roadway network. Each unique TMC code is used to identify a specific road segment. For example, in Table 3-5, TMC *114+0509* represents the WA-522 road segment with start location (*47.758321, -122.249705*) and end location (*47.753417, -122.277005*). However, that fact that WSDOT follows a linear referencing system based on mileposts poses challenges to matching the two different roadway layouts for data fusion.

Table 3-5 TMC Code Examples

TMC	Roadway	Direction	Intersection	Country	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	98115	47.753417,-122.27005	47.733752,-122.29253	1.608059

3.3 HERE Data

Similar to INRIX data, HERE combines data sources from multiple categories, including phone and auto GPS navigation devices. HERE data are collected separately from trucks and other vehicles, thereby making it possible to provide data for both trucks and passenger vehicles. HERE also adopted the Traffic Message Channel (TMC) as its base network, but the TMC network used by HERE is slightly different from that of INRIX. For each TMC, instead of providing speed data, HERE provides travel time data for passenger cars, trucks, and all vehicles (the weighted average travel time of both passenger cars and trucks); the travel time data are averaged into 5-minute bins. HERE data are currently available from September 2013 to December 2015 in the DRIVE Net system. The key information for HERE data is presented in Table 3-6.

Table 3-6 HERE Data Description

Columns	Data Type	Value Description
TMC	varchar	Unique ID for each segment-Traffic Message Channel (TMC) code
DATE	varchar	Date in format as mmddyyyy
EPOCH	varchar	Time of a day
Travel_TIME_ALL_VEHICLES	varchar	Average travel time of all vehicles
Travel_TIME_PASS	varchar	Average travel time of passenger cars

ENGER_VEHICLES		
Travel_TIME_FREIGHT_TRUCKS	varchar	Average travel time of trucks

3.4 WITS Data

Washington State's Incident Response (IR) Team collects and maintains traffic incident data in the Washington Incident Tracking System (WITS). WITS includes the majority of incidents that happen on freeways and Washington state highways, which totaled 550,376 as of March 2013. For each incident, the Washington State IR team logs details such as incident location, notified time, clear time, and closure lanes. The DRIVE Net team obtained the WITS datasets from 2002 to 2013 and integrated them into the DRIVE Net database. Several key columns are listed in Table 3-7.

Table 3-7 WITS data description

Columns	Data Type	Value Description
SR	varchar	State route ID, e.g., 005=Interstate 5
Direction	varchar	Route direction (NB=northbound, SB=southbound, WB=westbound, EB=eastbound)
MP	float	Milepost
Notifited_Time	datetime	The time when an incident was reported to the Incident Response (IR) program
Arrived_Time	datetime	The time when an IR truck arrived at the incident location
Clear_Time	datetime	The time when the incident had been fully cleared and all IR crews left the incident scene
Open_Time	datetime	The time when all lanes became open to the traffic and IR crews may still be on the incident scene

3.5 Weather Station Data

Weather data are retrieved from the National Oceanic and Atmospheric Administration (NOAA) weather stations in the region. The University of Washington Atmospheric Sciences Department hosts a website that records all the weather statistics from 209 weather stations in Washington state every hour. The DRIVE Net team developed a Java-based computer program to fetch the weather report in an automatic manner through an HTTP connection. The retrieved data are then imported into a database in the STAR Lab. The key information of the weather data is shown in Table 3-8.

Table 3-8 Weather data description

Columns	Data Type	Value Description
name	varchar	The weather station identifier
timestamp	datetime	24 hour time in integer format as YYYYMMDD hh:mm:ss
visibility	smallint	Visibility in miles
temp	smallint	Temperature in degrees Fahrenheit
dewtemp	smallint	Dewpoint temperature
wind_direction	smallint	Direction wind is coming from in degrees; from the south is 180
wind_speed	smallint	Wind speed in knots
pcpd	smallint	Total 6-hr precipitation at 00z, 06z, 12z and 18z; 3-hr total for other times. Amounts in hundredths of an inch.

Each weather station is associated with a latitude and longitude pair. In this case, weather data can be visualized on a mapping system.

3.6 Roadway Geometric Data

WSDOT's 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 an ESRI Shapefile are available to the general public, promoting data exchange and data sharing. Various roadway geometric data sets are available, including number of lanes, roadway widths, ramp locations, shoulder widths, and surface types. State route ID and locations marked by mileposts and accumulated mileage are also included in the WSDOT linear referencing systems. For DRIVE Net, these geometric data are stored in a spatial database for further processing. It is critical to connect roadway geometric data with traditional transportation data.

3.7 Ferry Data

Ferry data contain two parts: ferry location data and real-time ferry information data. WSDOT's 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 an ESRI Shapefile are available to the general public, promoting data exchange and data sharing. Ferry terminal data, both from public and private agencies, are contained in the "Ferry Terminal" data set. This data set is processed and stored in a GIS database of the DRIVE Net system. The real-time ferry information data, including ferry location, ferry speed, departing terminal, arriving terminal, available vehicle spaces, etc., are all queried from a WSDOT traveler information API.

3.8 Park and Ride Data

The static park and ride data also come from the GeoData Distribution Catalog online at http://www.wsdot.wa.gov/mapsdata/geodatacatalog/. The key information about park and ride lots within the data set includes name, county, city, address, zip code, number of spaces, and owner.

3.9 Transit Data

The public transit data are obtained from the OneBusAway API. For static data, bi-directional bus routes as well as stop information, including bus stop ID, bus stop address, latitude and longitude, are available in the data set; for dynamic data, real-time predicted arriving times for certain buses are available.

3.10 Bicycle and Pedestrian Count Data

Bicycle and pedestrian count data files were originally obtained from WSDOT in the xml format and stored in separate folders named by exact date. The DRIVE Net team converts and conflates these data into an SQL database. These data are collected from ten spots at ten different trails in Washington state. In the data set, basically, there are time interval information and corresponding bicycle/pedestrian counts in different directions. Time intervals also vary from 15 minutes to 1 hour, depending on the data collection year and trails. The collecting places' latitudes and longitudes are also available in the data set.

3.11 Car2go Data

The nationwide Car2Go data are available in the DRIVE Net system. These are obtained from the Car2Go API. All the real-time available Car2Go vehicle information can be obtained from the API, including vehicle location, vehicle name, engine type, and fuel level ranging from 0 (empty) to 100 (full). However, a Car2Go vehicle in use is not available from the API.

3.12 Interstate Freeway Elevation Data

The nationwide interstate freeway elevation data were originally extracted from the Google Earth API in an FHWA-funded project. For each freeway, the elevation data were for obtained every 10 feet. The automatic extraction process was developed by the UW STAR Lab research team. However, the original extracted data contained many outliers. The current data stored in the DRIVE Net database were further processed by using sophisticated filtering algorithms. In the database, elevation data for each state are stored as a table. The key information of the freeway elevation data is shown in Table 3-9.

Table 3-9 Freeway elevation data description

Columns	Data Type	Value Description
State	varchar	The state name
Route_name	Varchar	The interstate freeway name
Route_ID	Varchar	The ID of the route, which is unique

Direction	Varchar	The direction of the route
Longitude	Float	The longitude of the spot
Latitude	Float	The latitude of the spot
Milepost	Float	The milepost of the spot
Elevation	Float	The elevation value of the spot

Chapter 4 Data Quality Control

Although WSDOT provides a preliminary data quality assurance procedure to flag erroneous loop data, this procedure is still unable to capture other possible errors, such as loop detector sensitivity issues (Corey et al., 2011). Because of the environmental changes around loop detectors over time, the actual detection zone of these loops may increase or decrease, and these changes will consequently affect the accuracy of speed calculations. In fact, Zhang et al. (2003) stated that approximately 80 percent of WSDOT dual-loops suffer from severe sensitivity problems. Therefore, it is critically important to detect and correct possible loop errors before freeway performance is measured. A detailed loop data quality control mechanism is discussed in this chapter.

4.1 Introduction

This work is focused primarily on missing data imputation in transportation applications for the purpose of improving the coverage and accuracy of performance estimation. It is readily apparent that, as the rate of missing data increases as a result of detector malfunction or removal during quality control processing, the way that the missing data are handled quickly becomes the controlling factor in overall data quality. In this work, the issue was addressed by first developing algorithms to identify the various types of errors present in loop detector data in Washington state. Next, an imputation methodology was developed that is accurate and efficient both in terms of computational complexity and analyst time investment.

There are three principle steps involved in the DRIVE Net loop detector data quality process. First, a rule-based error detection algorithm that uses criteria developed to address common loop detector data errors is applied to flag erroneous and questionable observations. Second, a sensitivity adjustment algorithm is applied to detect loops with maladjusted sensitivity and a correction factor is applied to those deemed correctable. Finally, an imputation algorithm is applied to fill in missing observations and estimate prediction intervals. These steps are not strictly sequential, as the sensitivity adjustment step is completed in tandem with the error detection algorithm. All steps are applied to 20-second loop data on a monthly basis.

What follows in this chapter is a discussion of the causal factors related to missing and erroneous data, specifically with regard to the error detection algorithms and the assumptions inherent in the imputation procedure. Next, the approach used for missing data imputation is described. Finally, analysis completed to validate the data quality control and imputation procedures is described and discussed.

4.2 Background

For loop detector data, a loss of data becomes apparent in two ways. First, because of hardware malfunction or communications failure, a detector or cabinet may not report any data for a period of time. Second, data may be flagged as erroneous by a quality control routine and removed from the data set. It is likely that the unobserved values are not completely independent of all of the causal factors related to "missingness," and so a brief discussion is in order regarding missing data patterns and how they relate to the quantities of interest.

Most current work considers the occurrence of missing data under a probabilistic framework, with the pattern described by a statistical distribution (Rubin, 1976). The mechanism driving the missing data pattern is assumed to be ignorable if data are missing at random (MAR), which is only true when the distribution of missingness is not dependent on the unobserved or missing values themselves. That is, if the data set defined as X is constituted of both observed and unobserved components (X_{obs} and X_{mis}, respectively), then the probability that a value is missing depends only on X_{obs}, as shown in Equation 1 (Rubin, 1976; Schafer and Graham, 2002):

Equation 1: Probability of Missingness under MAR

$$\Pr(missing|X) = \Pr(missing|X_{obs}) \tag{4-1}$$

A simpler assumption would be that the missing data pattern is completely independent from both the observed and missing values, known as "missing completely at random" or MCAR. This assumption is implicit in the use of complete case analysis, where the missing observations are discarded and only the observed data are considered. Though this approach has been used in a great deal of transportation data analysis, the MCAR assumption is somewhat

difficult to justify and in some cases can result in certain locations and time periods being ignored altogether.

A number of factors affect the occurrence of missing and erroneous loop detector data. Some sources of missingness may be somehow related to the measured traffic parameters. For example, weather may affect both traffic speed and the occurrence of hardware failure. Likewise, if we consider time intervals during which no vehicles pass over a detector as "missing" in the sense that no speed measurement is made, lower volume time periods will likely be associated with higher occurrence of missingness. Such cases can be classified as MAR, and data from adjacent detectors can be used to estimate the missing values. The most prevalent influencing factors are related to hardware and communications malfunction or failure unrelated to the traffic parameters of interest. That said, if significant sequential blocks of data are missing, then making the assumption that the distribution of the missing data does not differ significantly from that of the observed data becomes less tenable. For this reason, it is important to develop imputation models individually for time periods that can be assumed relatively homogenous in terms of traffic behavior (i.e., monthly) and to flag those detectors that report little useable data during a given imputation interval.

One possible violation of the MAR assumption may be found in the quality control routines used to check the data for erroneous observations, a process based in part on the reasonableness of the observed values. Because of this, quality checking routines may flag some extreme (but still valid) observations for removal. With careful algorithm design, the occurrence and influence of such cases is expected to be minimal. The most common loop detector errors are discussed below, specifically in relation to the MAR assumption and potential violations thereof.

4.3 Causal Factors Related to Missing and Erroneous Data

4.3.1 Segmentation Error

Segmentation is caused when a vehicle detection occurs at the divide between two subsequent time intervals. When this occurs, the vehicle is counted during one interval, but the presence or occupancy is divided between the time periods. This may result in an unrealistically small

occupancy and a very high speed being computed from the observation. In reality, the true occupancy has little to no relation to the measured occupancy. For example, an observation with a volume/occupancy ratio of over 120 will almost certainly be flagged as an error. However, the MAR assumption only requires that the "missingness" not be attributable to unobserved values themselves. The estimated speed, then, is an artifact from a random event (i.e., a vehicle crossing the detector at a particular time) and can be safely removed without violating the MAR assumption. While it is true that the ability to detect such errors is somewhat dependent on traffic conditions, this mechanism can be described in part by neighboring detector observations. For all cases, it is important to set threshold values such that only those observations that are truly erroneous are removed (i.e., all values that can be considered plausible are retained).

4.3.2 Cross Talk

Cross talk occurs when two neighboring detectors interact, usually as a result of interference or short circuiting between the cables. Cross talk will usually result in very short and intermittent occupancy values, which can occur even when no vehicles cross over the detector of interest. Again, removing very low occupancy values on the basis of value thresholds has nothing to do with the actual occupancy at the location of the detector. Instead, as in the case of segmentation, the true values are unobserved. However, the MAR assumption is in this case questionable, because some recurring (i.e., long-term) hardware issue is causing the values to be removed. Similar to segmentation error, cross talk is often more easily detected in low traffic conditions, which increases the likelihood of removal. Again, the MAR assumption can be made more plausible by including a sufficient number of nearby detectors in the predictor set. Depending on the rate of occurrence, detectors prone to crosstalk may be removed entirely from the data set. For those retained, cross talk should be an exceedingly rare occurrence, and data can be assumed to be MAR if an adequate predictor set is used.

4.3.3 Stuck On or Off

If a detector is stuck on or off, the result will be a time interval (>> 20 seconds) during which volume and occupancy do not change. Typically, this error type is detected by setting a daily entropy threshold (Chen et al., 2001) under which a full day of data is removed. In this case, if

the removed block of data is significantly different from the observed data, then the MAR assumption may not be tenable. For this reason, the sensitivity of the employed imputation algorithm to such missing patterns should be investigated. Intuitively, it can be assumed that the accuracy of the imputation algorithm under this scenario will depend on the extent to which the missing data are representative of a typical day of operation. In any case, data imputed under such a scenario should be flagged as such to enable them to be excluded in subsequent analyses if needed.

4.3.4 Communications Failure

Communications failures will result in no data being recorded for a time interval. This may occur at the individual detector level or at the cabinet level, and may be caused by a variety of factors (see Rajagopal and Varaiya, 2007). If this occurs for a single time interval and is not due to some consistent underlying hardware problem, then the MAR assumption is plausible. However, some detector cabinets have a greater tendency for communications failure, which is indicative of underlying hardware issues. For relatively sparse and isolated failures, the MAR assumption is defensible. For extended time periods (i.e., >> 1 consecutive reporting interval or 20 seconds) or in cases with frequent communications loss, the MAR becomes less plausible as the length of the missing interval increases. Therefore, detectors that report a consistently elevated missing data rate should be flagged as suspect (which is typically done regardless of imputation method, see Chen et al., 2001). All imputed values produced in such a scenario should be flagged as such for possible exclusion in subsequent analyses, depending on the requirements of the analysis.

4.3.5 Sensitivity and Detector Health Issues

The sensitivity of a detector systematically affects the measured occupancy. For example, if the sensitivity is too high, vehicles will be detected before they reach the detector, and the detection will remain active for a brief time after the vehicle has passed. As a result, an unrealistically high occupancy will be recorded, leading to a lower speed estimate. In this case, an unrealistically high occupancy value may just represent a particularly high value instead of a random error. If such a value is removed on the basis of an occupancy threshold, then higher occupancy values will be removed with greater frequency than lower occupancy values. Therefore, it is critically

important to distinguish this scenario from the random error types in order to avoid violating the MAR assumption. One solution is to apply a sensitivity adjustment before error detection is performed to ensure that only truly erroneous occupancy values are removed.

For sensitivity or other detector health reasons, an entire day, week, or month of data is often discarded from the data set. This results in a non-probability sampling mechanism, as for a block of sampling intervals the probability of inclusion is zero (regardless of the actual unmeasured values). In this case, as previously mentioned, the reliability of the imputations depends to a large degree on the extent to which the missing data follows a distribution similar to that of the neighboring time periods. Of course, because the data are not observed, this cannot be assumed to be true, and some bias will likely be introduced. Again, data imputed under this scenario should be flagged as such for possible exclusion in subsequent analysis.

4.3.6 Other Missing Data Mechanisms

Additional causes of missingness that could violate the MAR assumption include construction activity, weather, and ongoing detector hardware issues. Therefore, the algorithm used to detect and eliminate erroneous values must have some mechanism for identifying the error type in order to make the distinction between random and not random missing patterns. In any case, as noted by Schafer (2010), standard ignorable missing data procedures are superior to ad hoc solutions, as the bias that can be explained by the observed values is removed, which is not true in general for ad hoc procedures.

4.4 Error Detection

The error detection algorithm is applied to each detector in turn on a monthly basis. The steps in the algorithm are based on criteria developed to address a number of common loop detector errors, including those described in the previous subsection. Figure 4-1 shows an overview of the hardware errors that are detected in the algorithm and links them to the methods used to address them in DRIVE Net.

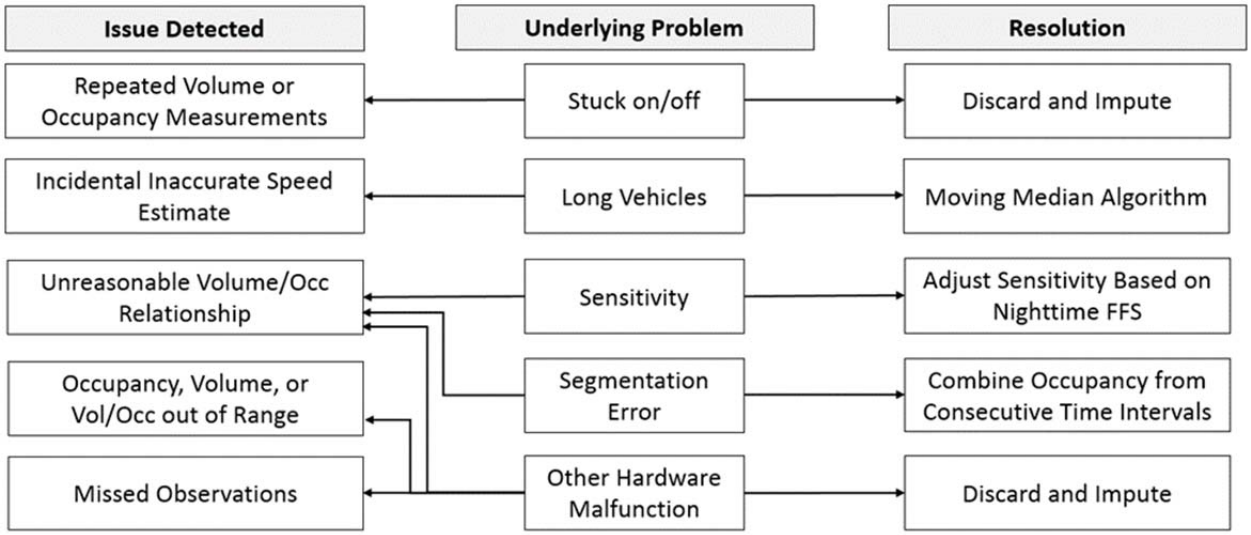


Figure 4-1 Common loop detector data quality issues and DRIVE Net solutions

The individual steps of the error checking algorithm are listed here, along with brief descriptions.

- Occupancy Split: During low volume time periods, the occurrence of segmentation can be identified as an interval with zero volume and non-zero occupancy. When this occurs, the "orphan" occupancy is added to that of the adjacent time interval, and the occupancy for the zero-volume interval is set to zero. Though this does not address segmentation error that occurs during higher volume time intervals, the influence of segmentation is less significant in those cases and is addressed to some extent by the moving median algorithm (below).
- Moving Standard Deviation: Detectors stuck on or off can be identified by computing the standard deviation of the volume/occupancy ratio on a moving window. Any points falling into a window with a sample standard deviation below a fixed threshold will be flagged as errors.
- Daytime Missing Data: If volume or occupancy is zero during the time interval from 5:00AM to 10:00PM, it is flagged as such. Though this does not necessarily indicate an error, it is flagged so that the number of zero observations in the daytime can be compiled in the output statistics. This way, detectors with an abnormally high occurrence of zero observations can be identified. Note that a zero volume time interval in most cases should

not be considered missing for the purpose of imputation, but the volume/occupancy ratio for the same interval should be considered missing and imputed.

- Sensitivity Adjustment: The sensitivity adjustment is only applied to detectors on the mainline. The late night free flow speed (FFS) is computed for each detector and is compared to the expected value for the location (65 mph for most freeway segments in western Washington). If the observed late night free flow speed is within 10 percent of the expected value, no adjustment is applied. If it is within 10 percent to 50 percent of the expected value, then the adjustment factor shown below is applied to the observed volume/occupancy ratio. If the late night FFS is not within 50 percent of the expected value, then the detector is flagged as questionable in the monthly sensitivity table.

$$adjustment\ factor = \frac{expected\ FFS}{late\ night\ FFS} \tag{4-2}$$

- Parameter Thresholds: Thresholds are applied to volume and the volume/occupancy ratio. Specifically, volume/occupancy values above 120 are flagged as erroneous (corresponding to a speed of greater than 90 mph), and volumes above 22 vehicles/20 seconds are flagged as erroneous. Most often, unrealistically high volume observations are the result of chattering, while unrealistically high volume/occupancy values could be the result of one of several hardware malfunctions. Note that this step is applied after the sensitivity adjustment to ensure that only truly erroneous values are removed by the parameter thresholds.
- Moving Median Algorithm: A 2-minute moving median algorithm is applied to the volume/occupancy ratio to reduce the influence of long vehicles. The equation used by WSDOT to compute traffic speed from single loop detector observations assumes a constant vehicle length, and so the time intervals in which long vehicles pass over the detector often produce unrealistically low speed estimates.
- Nighttime Missing Values: Missing volume and volume/occupancy values observed during the late night (11:00PM to 4:00AM) are flagged as missing. Note that, in most cases, this does not indicate a hardware malfunction or other error. The flags are applied for record keeping purposes.

- Aggregate and Report Results: In this final step, the various error and missing flags are summarized in a daily detector health table. The sensitivity information is similarly recorded in a monthly sensitivity table, including flags indicating whether and to what extent the volume/occupancy ratio was adjusted. If no data are returned for a given month, a record is made in the sensitivity table indicating no data for that month and detector.

Figure 4-2 (below) summarizes the error detection and sensitivity adjustment algorithm.

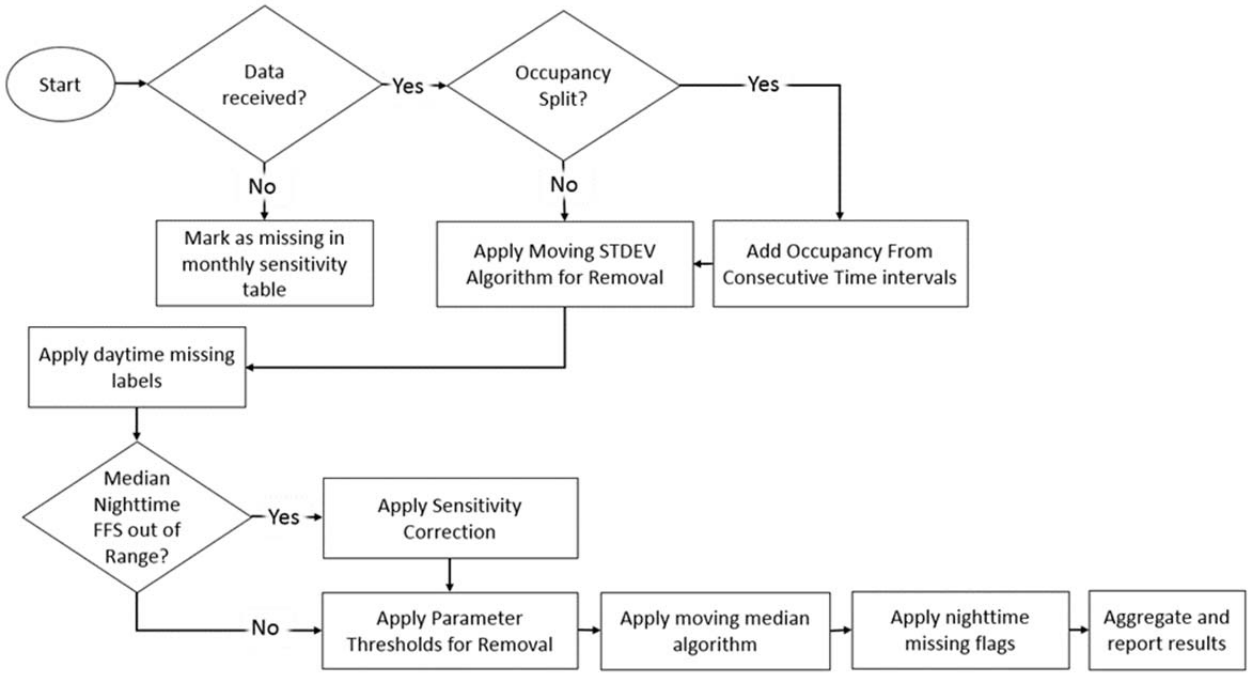


Figure 4-2 Error detection and sensitivity correction algorithm

4.5 Imputation Approach

Substantial previous research has focused on imputing missing traffic sensor data. One widely used method, described in Chen et al. (2001) and Chen et al. (2003), applies an iterative, pairwise linear regression that uses nearby detector observations as predictors to predict missing loop detector observations. Essentially, for each detector of interest, individual linear regression models are developed for each nearby detector (both adjacent and upstream/downstream), and the missing value is estimated as the median of predictions from all detectors reporting useable

data. In this and similar approaches, no method is given for estimating the uncertainty in the predicted values. In addition, the accuracy of this approach degrades during congested conditions. Although some methods have applied more complex models with second order and interaction terms, the process of automating the model building, predictor selection, and model evaluation on a very large scale (billions or rows of historical data) would become quite complicated. This illustrates an obvious problem with using parametric models of increasing complexity to describe the relationships between neighboring detectors, as the relationships vary by geometry, time, and traffic conditions. A number of methods have been developed to apply more robust, multiple imputation models (e.g., Henrickson et al., 2015; Ni et al., 2005a; Ni et al., 2005b), as well as time/space tensor models (Tan et al., 2013) and several dimension reduction approaches (Qu et al., 2009; Li et al., 2013). However, although such methods may provide a high level of accuracy, they are not computationally tractable for the large loop data sets used in DRIVE Net.

To address the imputation challenge presented by the massive and heterogeneous loop detector dataset in the DRIVE Net system, four objectives were identified:

1. The method must provide reasonably accurate and un-biased replacement values for use in engineering analysis.
2. The method must be computationally tractable for very large data sets.
3. Because the imputation process will be largely automated, the method must not require any strict assumptions regarding the statistical properties of the data and must be relatively insensitive to model specification.
4. The approach must provide some method of accurately estimating the uncertainty in the estimated values.

On the basis of these requirements, several tree-based ensemble methods were investigated as possible solutions, as were fast approximate nearest neighbors methods.

4.5.1 Random Forest Regression

Non-parametric, decision tree-based classification and regression is a family of machine learning methods that have been applied in a variety of fields in the last decade. To build a regression

tree, the predictor space is split into J (possibly high dimensional) non-overlapping, rectangular regions, such that the total sum of squares of the response is minimized. This process is typically completed by using greedy recursive binary splitting such that, at each stage, the split resulting in the greatest reduction in total sum of squares is selected (Hastie *et al*., 2009). When this partitioning is completed, a tree-like structure emerges in which each terminal node or "leaf" contains an increasingly homogeneous subset of observations. Predictions can then be generated by assigning an observation to a single leaf and obtaining the mean observed value for that leaf.

This single regression tree approach is somewhat prone to overfitting, and building an optimal tree may be difficult because of the greedy splitting algorithm. Although these limitations can be addressed to some extent by controlling the size of the tree and through careful predictor selection, a number of methods have been introduced to improve the predictive power and robustness of tree-based classification and regression models. Introduced by Breiman (2001), Random Forest is an ensemble approach that builds a "forest," or set of trees, each using a bootstrap sample of the available data. For each tree, samples are drawn with replacement from the available data, and a regression tree is built with this subset. Instead of considering all predictors in identifying the best split at a given stage, with this approach, a random sample of the available predictors is considered to include some randomness in the individual trees. Thus, though individual trees may be suboptimal and/or over-fit, excellent predictive accuracy can be achieved by aggregating over the predictions from all trees. Predictions are produced by assigning an observation to a terminal leaf in each tree in the forest and obtaining the (often weighted) average of all observations in the assigned leaf(s). The benefit of the ensemble method is that, by introducing some randomness in the individual predictive models, the resulting ensemble is robust to noise and overfitting, provided that a sufficient number of trees is used (Liaw and Wiener, 2002). Furthermore, predictive accuracy can be closely estimated by applying "out of bag" error estimates to the training data. That is, because only a subset of the training data is used to build each tree, a reasonable estimate of test error can be produced by generating predictions for each training observation that are based on only the trees that did not include the observation. In addition, because individual tree construction is quite fast and trees can be built in parallel on multi-core computing hardware, model training is both fast and scalable.

Meinshausen (2006) introduced the quantile regression forest, which enables the statistical properties of the predictions (rather than a simple mean prediction) to be estimated. In essence, this approach builds a random forest as described above, but it then generates predictions by 1) assigning an observation to the appropriate terminal leaf on each tree in the forest; 2) gathering all observations from the assigned terminal leaf on each tree (rather than the mean or mode); and 3) estimating the prediction quantiles from these observations. The method applied in this work is based on the method described in Meinshausen (2006), except that each tree is fully developed (such that each terminal leaf contains a single observation) and the quantiles are estimated from those observations. The method was adjusted to take advantage of the efficient computational tools available in the python sci-kit learn libraries, which do not support the methods described in Meinshausen (2006) in their original form. With this methodology, the mean and prediction intervals (e.g., 95 percent) for each missing value can be computed. Random forest models have been used previously for imputation, for example Stekhoven and Bühlmann (2012) described an iterative procedure somewhat similar to the procedure of multiple imputation by chained equations introduced in Buuren and Oudshoorn (1999). However, the process used here is a non-iterative, single imputation approach with the following desirable characteristics:

1. Efficient – Unlike many more sophisticated methods, this approach is computationally tractable for very large data sets.
2. Robust – By including a relatively large number of trees in the model, the risk of overfitting or missing important features in the data is minimized.
3. Provides accurate uncertainty measures – The simplified quantile regression forest approach can generate both mean predictions and reasonably accurate prediction intervals.
4. Requires few assumptions about the structure or statistical properties of the data – In the proposed approach, few assumptions are made about the distribution of underlying data or the nature of the relationship between predictors and response.

The algorithm applied here can be summarized by the following steps:

Loosely following the notation in Stekhoven & Bühlmann (2012), for each month, let each detector s produce the vector of observations $y^{(s)}$, which consists of both observed ($y_{obs}^{(s)}$) and missing ($y_{miss}^{(s)}$) components. Let X indicate the $n \times p$ matrix of predictors $X_1, X_2, \ldots X_p$, where X_j indicates the vector of observations from an adjacent, upstream, or downstream detector. The model for detector s is then fit by using only $y_{obs}^{(s)}$ and the rows in X corresponding to $y_{obs}^{(s)}$, denoted $X_{obs}^{(s)}$ as follows:

$$y_{obs}^{(s)} \sim X_{obs}^{(s)} \tag{4-3}$$

Predictions and prediction intervals for $y_{miss}^{(s)}$ can then be drawn by using the fitted model and the rows in X corresponding to $y_{miss}^{(s)}$, denoted $X_{miss}^{(s)}$:

$$y_{miss}^{(s)} \sim X_{miss}^{(s)} \tag{4-4}$$

Note that, in most cases, missing values will be present in most of the predictors contained in both $X_{obs}^{(s)}$ and $X_{miss}^{(s)}$. To overcome this, all missing values in each predictor are filled in with the mean value for that predictor. This approach is often applied in random forest models, and although there is some loss in predictive power and interpretability, it can provide good results under reasonable missing rates when a sufficient number of regression trees is used.

4.5.2 Aggregation Levels

The majority of current research on the topic of imputing missing traffic sensor data relies on data aggregated to 1-minute, 5-minute, or longer time intervals. This reduces the impacts of random noise and results in better spatial and temporal correlation structures. In Washington state, much of the performance reporting is based on 5-minute intervals, which further strengthens the argument for imputing pre-aggregated data. However, this research follows the principle of imputation before aggregation for several reasons. First, for data that are measured at the 20- or 30-second level, some elementary imputation is implicitly applied in the aggregation step. Thus, many "complete" 5-minute intervals are based on incomplete data, and many 5-minute intervals are marked as missing even when not all of the contributing 20-second values are actually missing. Second, by applying a principled imputation method at the lowest available

aggregation level and aggregating the complete data set to the desired time intervals, better performance can result even if the per-observation accuracy of imputation is lower than what could be achieved by imputing pre-aggregated data (Ni, et al. 2005a; Ni, et al. 2005b).

4.6 Validation

The validation work presented here is for a set of loop detector cabinets on I-5 between mileposts 150 and 160. In total, this included 400 loop detectors in both travel directions and with varying sample rates (depending on the quantity of useable data reported by each loop detector). Slightly under 2.52 million individual observations were made during the month of June 2012. By setting observations missing at random and applying the proposed imputation methodology, a reasonable estimate of the performance of the algorithm could be generated. The data used for this analysis included detectors with a range of preexisting missing and erroneous observations; a summary is provided in Table 4-1. Note that the relative contribution to the overall statistics was lower for detectors with higher rates of missing data, as a smaller set of data was available for testing. However, in the analysis presented below, detector-wise performance measures are provided.

Table 4-1: Summary of missing rates over all detectors used in validation

	Missing Rate
Max	*0.986*
Min	*0.006*
Mean	*0.229*
Standard Dev.	*0.228*

The measures used to assess the accuracy of the imputed values and coverage of the prediction intervals included mean absolute percentage error (MAPE), mean percentage error (MPE), and the percentage of observations falling within the 95 percent prediction intervals, or coverage level (CVL). The formulas for these measures are given below.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{T_i - I_i}{T_i}\right| \quad (4\text{-}5)$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i - I_i}{T_i} \quad (4\text{-}6)$$

$$CVL = \frac{1}{n}\sum_{i=1}^{n} 1_{\{L_i, U_i\}}(T_i) \quad (4\text{-}7)$$

where

n = the total number of observations used to compute MAPE

T_i = True (observed) value for index i

I_i = Imputed value for index i

L_i, U_i = Lower and Upper prediction interval bounds, respectively

$1_{\{A,B\}}(Y)$ = Indicator function, 0 or 1 depending on whether or not $A \leq Y \leq B$

Table 4-2 provides a summary of the results, with a mean absolute percentage error of approximately 6.6 percent and an average error (a measure of bias) of approximately -0.124 percent. The overall prediction interval coverage was approximately 91.7 percent, which was somewhat less than the expected 95 percent. Note however that this analysis included some detectors that reported very little useable data and in fact would likely have been excluded from any analysis on this basis. A more thorough analysis of the error and coverage is given below.

Table 4-2 Summary of imputation results

Observation count	Overall MAPE	Overall MPE	Overall CVL
2,519,973	6.66%	-0.124%	91.7%

Figure 4-3 shows a histogram of the MAPE for all observations. Note that the majority (more than 85 percent) fell below 11 percent at the 20-second level, which is competitive with most published work on the topic (e.g., Al-Deek et al. 2004).

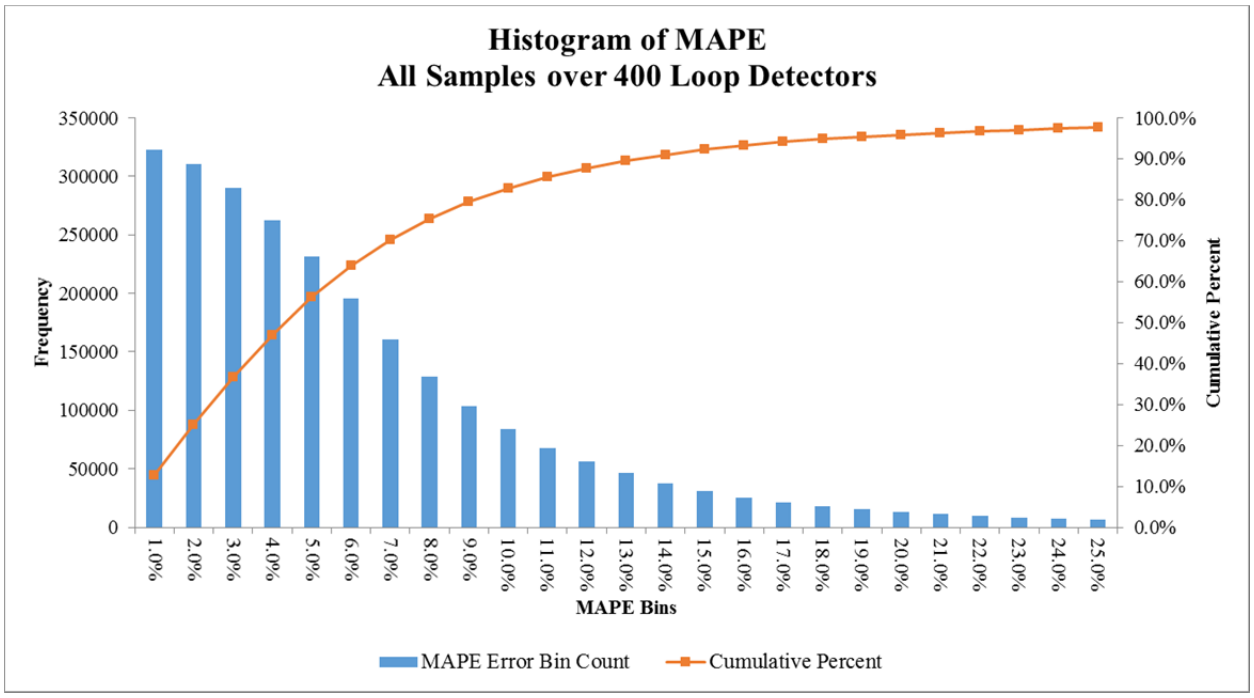


Figure 4-3 Histogram of MAPE for all observations

Figure 4-4 shows a histogram of the MAPE by detector, and again the majority (over 90 percent) of detectors reported a MAPE of less than 11 percent. As discussed in Ni et al. (2005) and Henrickson (2014), with minimal bias this level of accuracy at the 20-second level will likely improve significantly when the observations are aggregated to the 5-minute level for use in analysis. Figure 4-5 shows a histogram of the percentage error over all observations and loop detectors.

Figure 4-6 shows a plot of the observed vs. imputed values for a single loop detector. The line is plotted at 45 degrees and illustrates what would be considered a perfect fit to the observed values. This particular detector was chosen to illustrate the fit because the randomly sampled values included a wide variety of traffic conditions and yet still demonstrated that the proposed methodology is both accurate and highly responsive to changing traffic conditions. Figure 4-7 shows a sample of the imputation results. Note that while the observations are ordered by time, this is not a continuous time series because the values are randomly sampled from the available data. The error bars show the bounds of the 95 percent prediction interval

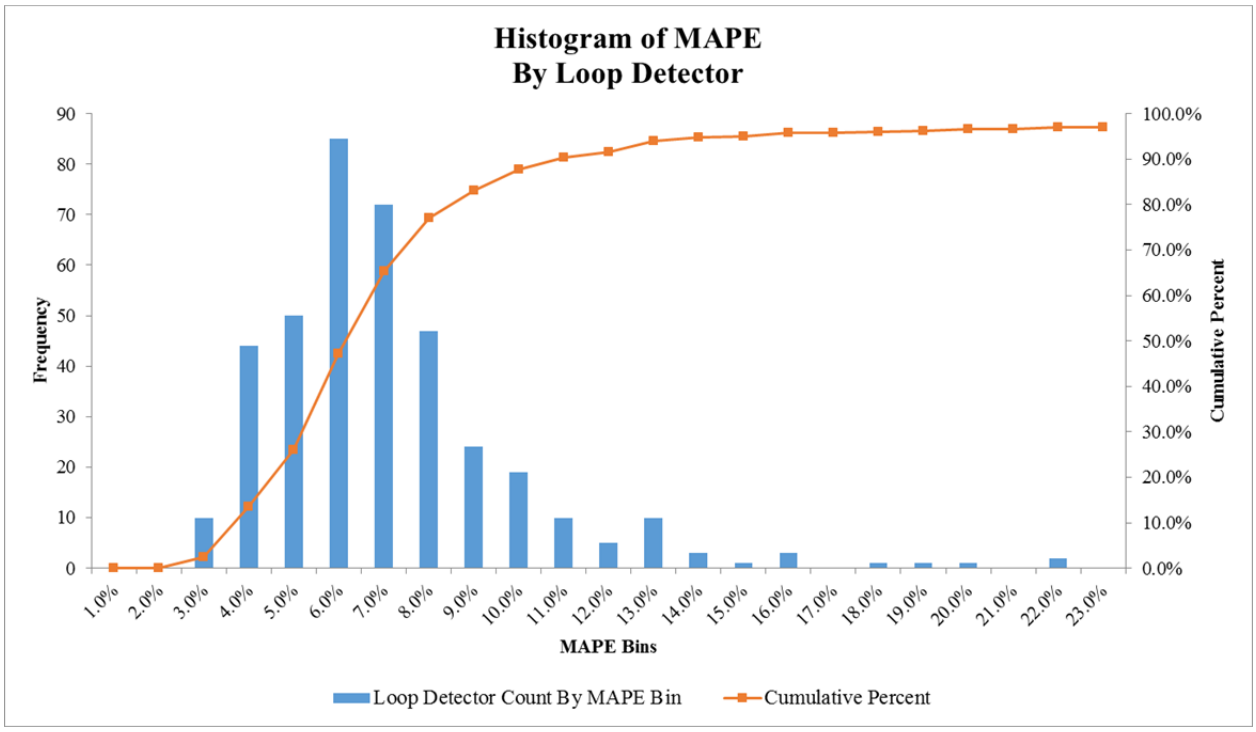


Figure 4-4 Histogram of MAPE by loop detector

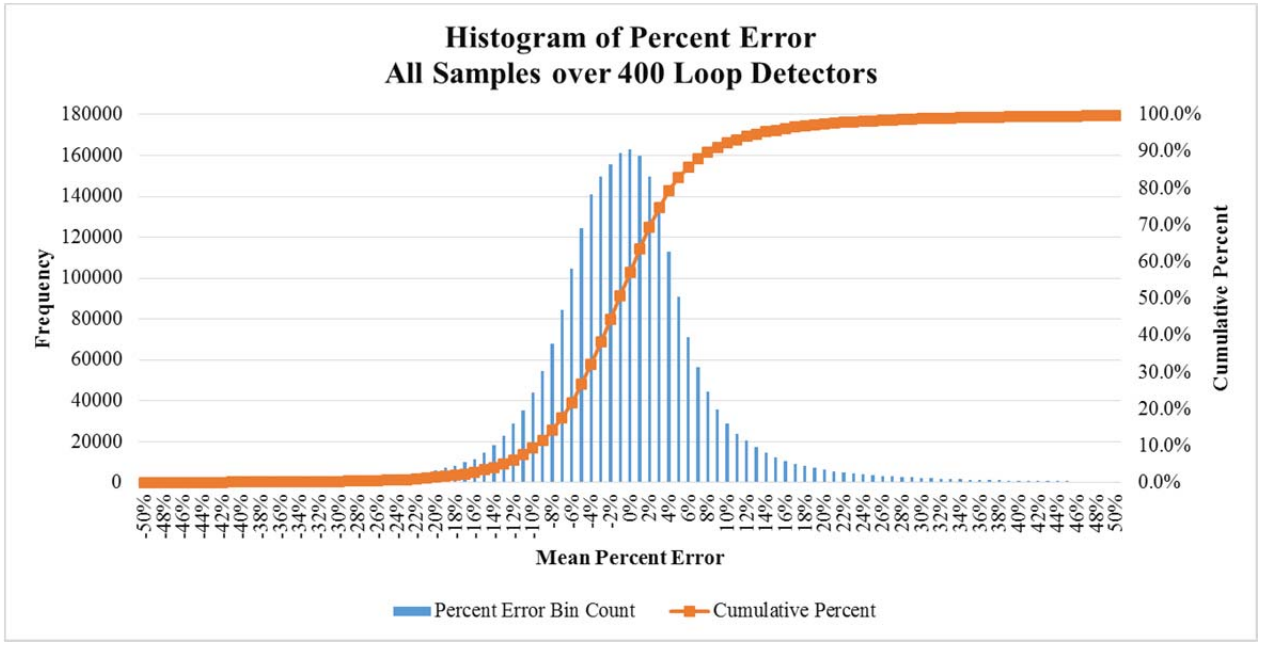


Figure 4-5 Histogram of percentage error

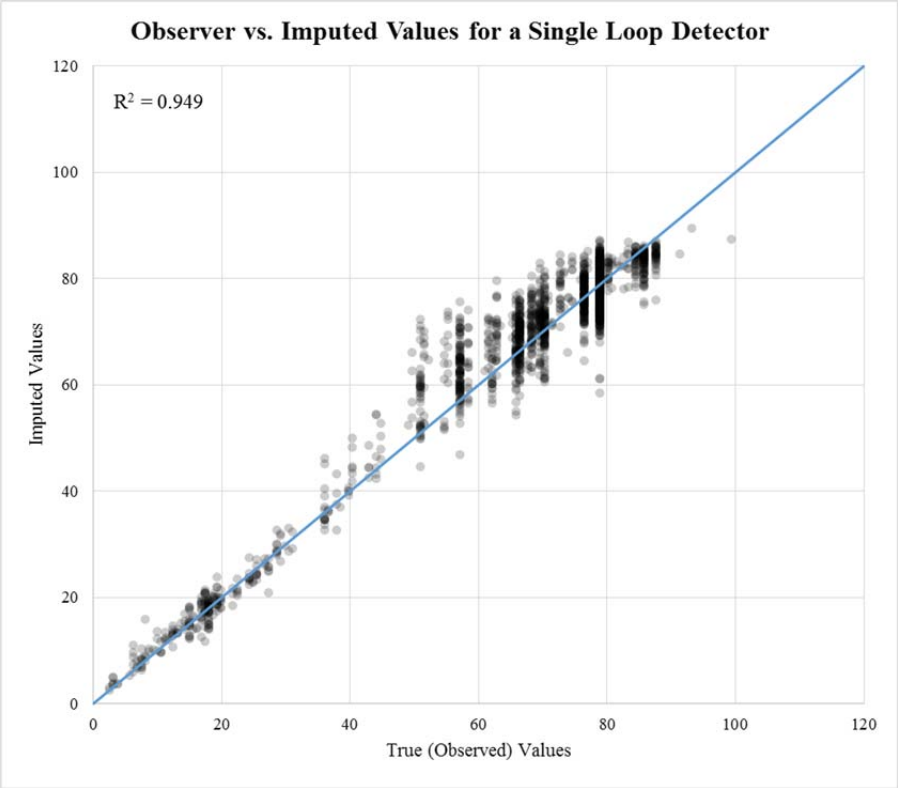


Figure 4-6 Observed vs. imputed values for a single detector (~2000 observations)

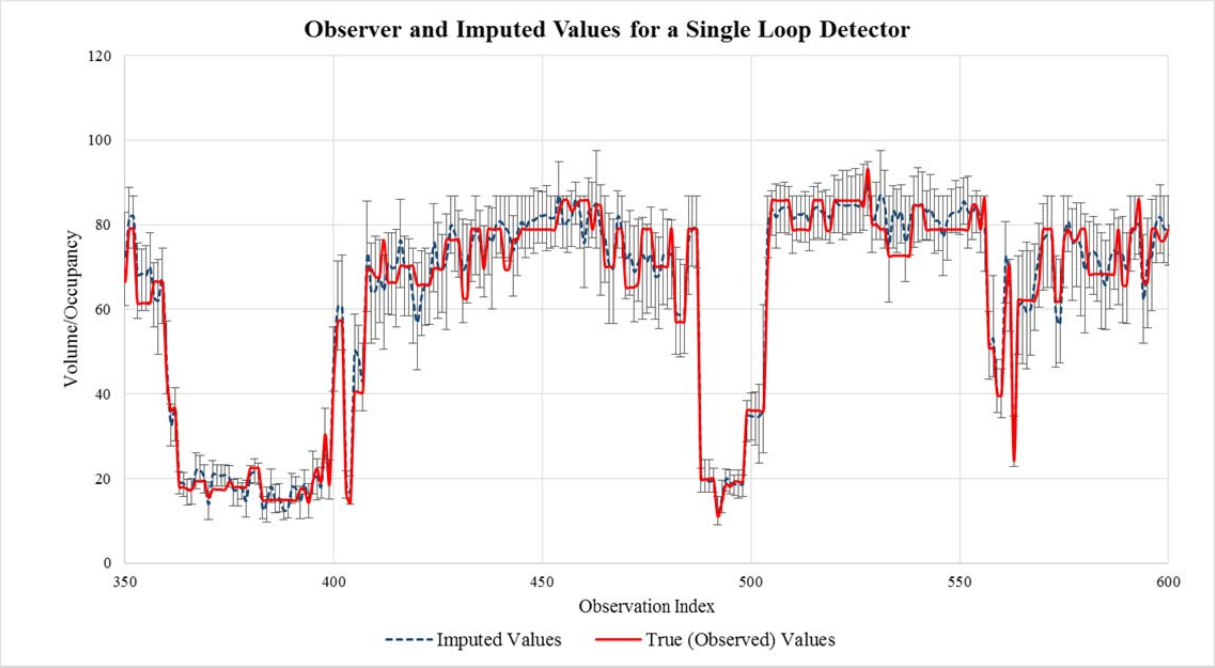


Figure 4-7 Sample of imputation results

Chapter 5 DRIVE Net 4.0: System Design and Implementation

After many years of development, DRIVE Net already had significant capabilities for data handling, transportation analysis, and traffic data visualization. Nevertheless, new features were added to DRIVE Net 4.0, such as security access control and database connection optimization. These new features enhance the security and performance of the system. Specially, because of the larger number of data sources merged into DRIVE Net—for example, pedestrian and bicycle, public transit, park and ride, Car2Go, and ferry data—DRIVE Net 4.0 has more data visualization and analysis modules and corresponding capabilities.

In DRIVE Net 4.0, several challenging problems that remained unsolved in DRIVE Net 3.0 have been addressed. The most critical issue, database interface mapping between the data sets and transportation analysis logic, has been improved. In addition, the key abilities in the previous version—such as geo-processing power to store, analyze, and manipulate geographic data with OpenLayers and OpenStreetMap (OpenLayers, 2013; OpenStreetMap, 2013)—have been enhanced, making the DRIVE Net system a more flexible yet reliable alternative web-mapping product.

5.1 System Functionality

According to the DRIVE Net phase II proposal and requirements of WSDOT, the fundamental data, visualization, and analysis functions of DRIVE Net are as follows:

5.1.1 Database Functionality

To support the current WSDOT requirements and future data analysis needs, the DRIVE Net system should supply the following database functionality:

- Data entry/update (either in real time or via batch updates) for all data sets that are part of DRIVE Net. This includes a process for updating data stored in the system, whether those data sets are obtained in real time from WSDOT systems (e.g., 20-second freeway data), computed from those data (e.g., 5-minute freeway data), or obtained in batch files from WSDOT (e.g., GIS base files, WITS, INRIX, HERE, and crash records).

- Data quality control for each data set. This includes performing data quality control on each data set and maintaining a record of the quality control outcomes.
- Data retrieval. This includes the ability to extract all base data stored in DRIVE Net. The data need to be available by date or date range, and either by specific location, corridor, or partial corridor.

5.1.2 Visualization Functionality

DRIVE Net should be a user-friendly system that supports interactions with users for data exchange, traffic reporting, and analysis tasks:

- Visualization of downloaded data
- Visualization of travel time analysis/HOV analyses
- Visualization of multi-modal analyses
- Visualization of safety analyses.

5.1.3 Analytical Functionality

The analytical functionality required for DRIVE Net is intended to meet the basic performance monitoring needs of WSDOT. The intent of the system is to provide these capabilities state-wide.

- Travel delay quantification and congestion analysis. This includes a universally accessible congestion diagnosis module to analyze congestion onset and evolution and a series of visualization modules for reporting and analytical requirements.
- Safety analysis. Crash models to estimate expected accident frequencies and identify key associated factors and a universally accessible regional map module to visualize both accident hotspots and safety improvements.
- Travel time analysis. The modules should be able to calculate travel times, throughput productivity measurements, etc. that are based on different data sources such as freeway loop, INRIX, and HERE.
- Multi-modal analysis. The capability to calculate and analyze different transportation modes such as bicycles, walking, park and rides, transit, and ferry service.

5.2 System Design

The new system adopts the "thin-client and fat server" architecture with three basic tiers of Web application: the presentation tier, logic tier, and data tier, as shown in Figure 5-1. The presentation tier includes the user interface terminal through which users interact with the application. The logic tier, which is also called the computational tier, is the core component of the DRIVE Net system. It performs computations to assist in customized analysis and decision making based on users' interactive input. The 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 the user interface and dealing with user interactions.

There are three reasons to adopt the thin-client architecture: First, no plug-in and installation are required at the client side except a basic browser, which ensures the highest level of compatibility. Given that the system is designed for customers with constrained network functions, minimal requirements on the client side are most desirable. Second, there are fewer security concerns since all the data and computational tasks are manipulated and performed on the server side, and the client is only responsible for user interaction and results presentation. Third, mature frameworks for building thin client Web applications could be re-used to boost development productivity. However, thin-client architecture does have its drawbacks. One major disadvantage is that the performance of the system depends solely on the server and, as a result, excessive user requests greatly affect system efficiency. This has become more manageable in recent years with the continuous advancement of cloud computing technologies such as Amazon Web Service, whose cloud servers are fully designed to improve system performance.

The data communication flows in the DRIVE Net system can 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 retrieves the related data information from the data warehouse.

3. The warehouse sends back the requested data, and the web server performs the computational tasks by using either the built-in analytical tools or external statistical modules provided by R Server.

4. If geospatial analysis is involved, the web server connects to the OpenStreetMap Server and requests the map.

5. Analysis results as well as the map are then returned to the client. The web browser displays the results or visualizes the returned objects on the map.

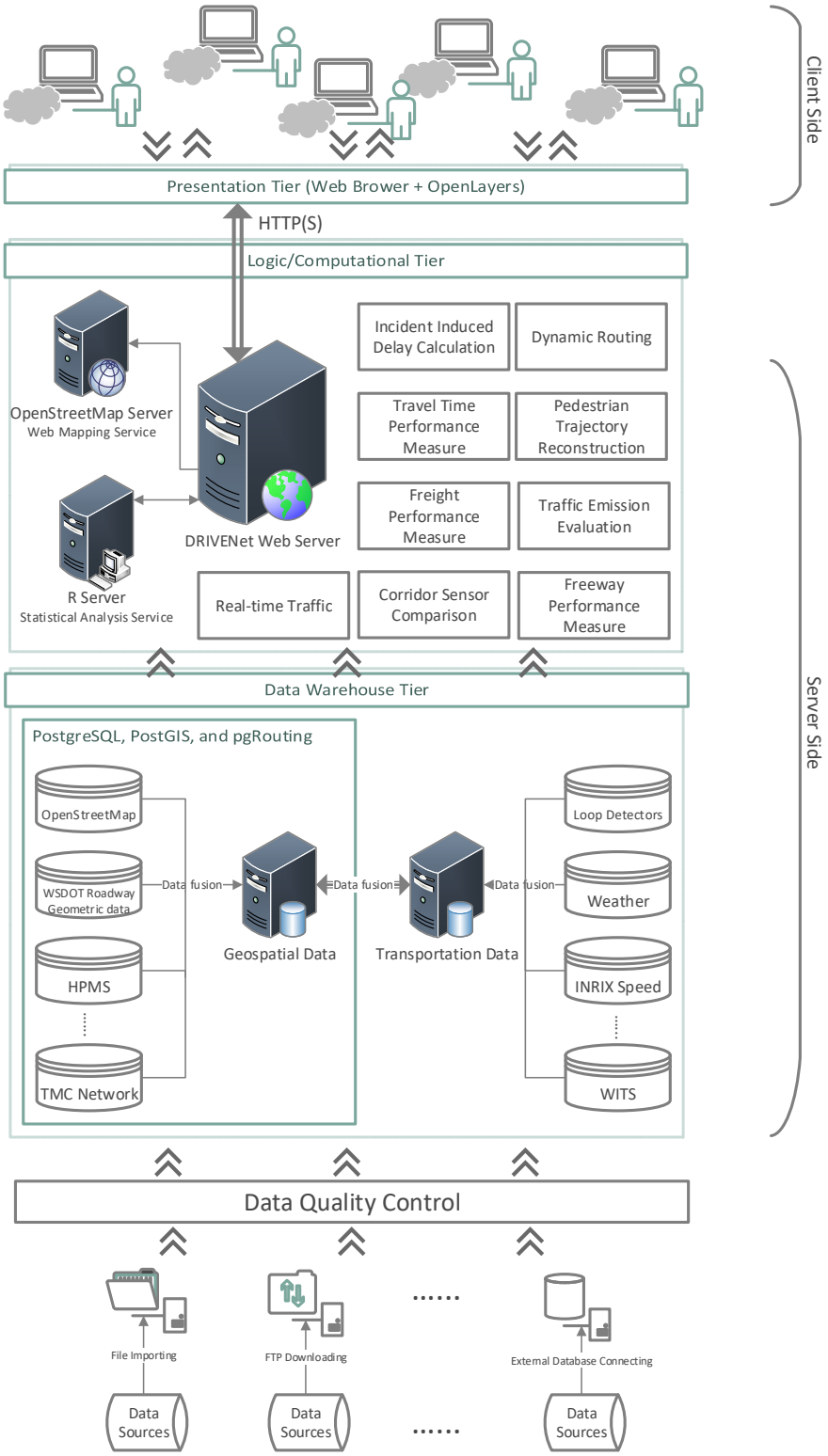


Figure 5-1 DRIVE Net 4.0 architecture

5.3 Geospatial Database Design

Because of the increasing amount of incoming data, multiple servers are configured to archive these data. To better balance computational resources and allow fast data access, transportation data and geospatial data are stored separately. The transportation data are managed by Microsoft SQL Server 2010, and all the databases are indexed and optimized on the basis of projected needs. However, the traditional method for handling geospatial data sets is to utilize commercial GIS software packages. Unfortunately, transportation agencies have to spend considerable amounts of time and financial resources purchasing and maintaining the software (Sun et al., 2011). In addition, because most commercial software is not designed as open architecture, transportation agencies have to provide the spatial data in strict accordance with the GIS file format of used by the commercial software. These restrictions incur inconveniences and reduce flexibility for both users and developers. Moreover, file-based data management systems have inherent disadvantages for processing tremendous amounts of data efficiently. Fortunately, the emergence of new geospatial database techniques can alleviate the burden of file-based geospatial data management and analysis. Similar to the traditional Relational Database Management System (RDBMS), geospatial databases can optimize the geospatial data management and analysis by using Structured Query Language (SQL) techniques and spatial indices. In addition, geospatial databases enable a variety of geo-processing operations that traditional relational, non-spatial databases cannot complete—for example, whether two polylines intersect, or whether points fall within a spatial area of interest. For this study, non-spatial relational databases were used to store traffic-related information such as loop detector data and INRIX data. This created a critical issue: how to best represent and manage the dynamic transportation data in a context of hybrid spatial and non-spatial databases. Because more and more location-aware transportation data are available for advancing Big Data initiatives, this issue is becoming more pressing.

For the new system, PostgreSQL with extender PostGIS and pgRouting were adopted to maintain geo-data and perform spatial modeling, as outlined in Figure 5-2. 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 the same jobs, open source

projects are generally more academic in nature, and commercial products usually have expensive license and usage restrictions. The rest of this section introduces more details about PostgreSQL, PostGIS, and pgRouting.

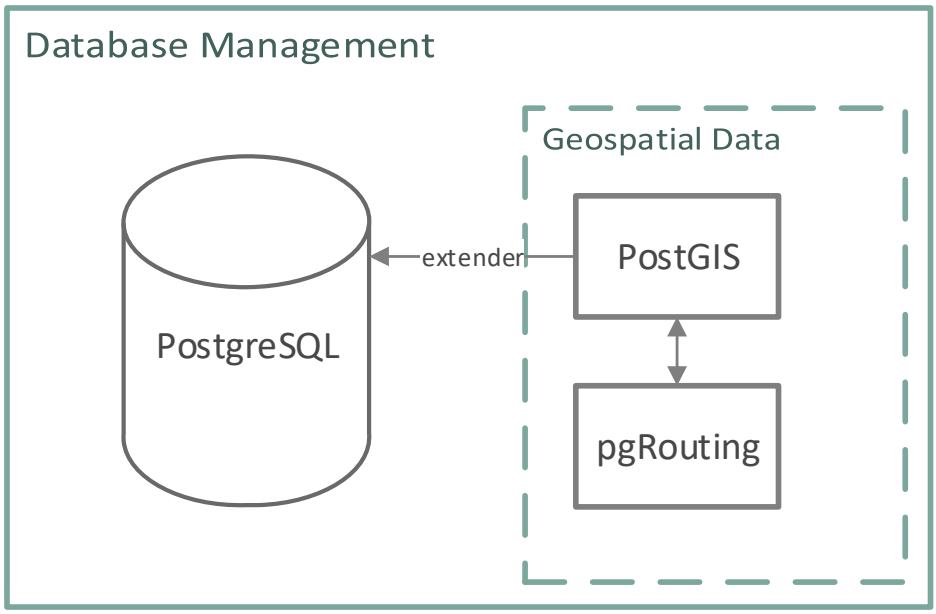


Figure 5-2 PostgreSQL, PostGIS, and pgRouting

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 it is able to compete against well-known commercial products such as Oracle, IBM DB2, and Microsoft SQL server. Some advanced and unique features that distinguish it from others include table inheritance, support for arrays, and multiple-column aggregate functions. Moreover, the active global community of developers continually updates PostgreSQL with the latest database technology.

PostGIS is a spatial database extender that builds on the PostgreSQL tabular database (Obe, 2011). The PostgreSQL/PostGIS combination offers support 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 a size of 10,000 feet:

Select ST_Buffer(the_geom, 10000) from county_polygon

pgRouting is an extension of PostGIS/PostgreSQL geospatial database that provides a set of routing-related SQL functions (pgRouting, 2013). pgRouting supports various routing algorithms, including shortest path Dijkstra (Dijkstra, 1959), shortest path A* (Hart et al., 1968), shortest path shooting*, traveling salesperson problems, and driving distance calculation. Meanwhile, its open source framework makes it convenient for developing and implementing user-specified routing algorithms. More advanced algorithms such as Multimodal Routing support, Two-Way A*, and time-dependent/dynamic shortest path will be included in the near future.

5.4 Key Techniques in System Implementation

As mentioned in the previous section, the DRIVE Net architecture was redesigned to meet challenges. To reduce costs and boost productivity, multiple open source products were utilized. Relying on open source products, the DRIVE Net team not only took advantage of code-sharing and collaboration with a broad community of developers but also contributed to open source projects. The core open source products combined into the DRIVE Net system are explained in the remainder of this section.

5.4.1 OpenStreetMap and OpenLayers

OpenStreetMap (OSM) is a collaborative project that has created 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, University College London established the OSM project in July 2004, and it is one of the most prominent and famous examples of Volunteered Geographic Information, a concept introduced by Goodchild (2007, 2008). The process of maintaining OSM data is termed “crowdsourcing” and is being used by a number of other commercial companies, such as Google and TomTom. In crowdsourcing, defined by Brabham as an “online and distributed problem-solving and production model,” labor-intensive tasks are

distributed to large groups of users, and this has allowed volunteers to create and update geospatial data on the Internet. By January 2013, OSM had over one million registered contributors and 20,000 active users worldwide, and the number continues to rise dramatically (Wood, 2013). Besides governments, OSM has obtained strong support from commercial companies. For instance, Yahoo Maps made its 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 (Microsoft, 2010).

One major reason for DRIVE Net to choose OSM is its low cost in comparison to commercial data sets, 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). Using OSM to replace Google Maps helps DRIVE Net avoid potential charges by Google, Inc in the future that might eventually prevent the project from growing. In addition, in keeping with the theme of eScience, the DRIVE Net developers prefer open source products over commercial ones because they can help share ideas, drive innovation, and boost productivity for the entire community.

High-resolution and qualitative geographic information such as that shown in Figure 5-3 makes OSM an appealing replacement for Google Maps. Recent research confirms the good quality of OSM and its ability to compete against commercial geodata, especially for urban areas. Zielstra and Hochmair (2011) used the commercial data sets NAVTEQ and TeleAtlas, as well as the freely available dataset TIGER/Line, to quantify the coverage of OSM in the United States. The results indicated that “there is strong heterogeneity of OpenStreetMap data for the U.S., in terms of its completeness.” A similar study was done in Germany by Zielstra and Zipf in 2010 (Zielstra and Zipf, 2010). The paper stated that some projects had already replaced proprietary data with rich OSM data in larger cities. In the U.K., Haklay (2010) compared OSM with the Ordinance Survey (OS) Meridian data set by evaluating the accuracy, completeness, and consistency of their positions and attributes. The analysis concluded that “OSM information can be fairly accurate,” with a positional accuracy of about 6 meters and an approximately 80 percent overlap of motorway objects in comparison to the OS data set.

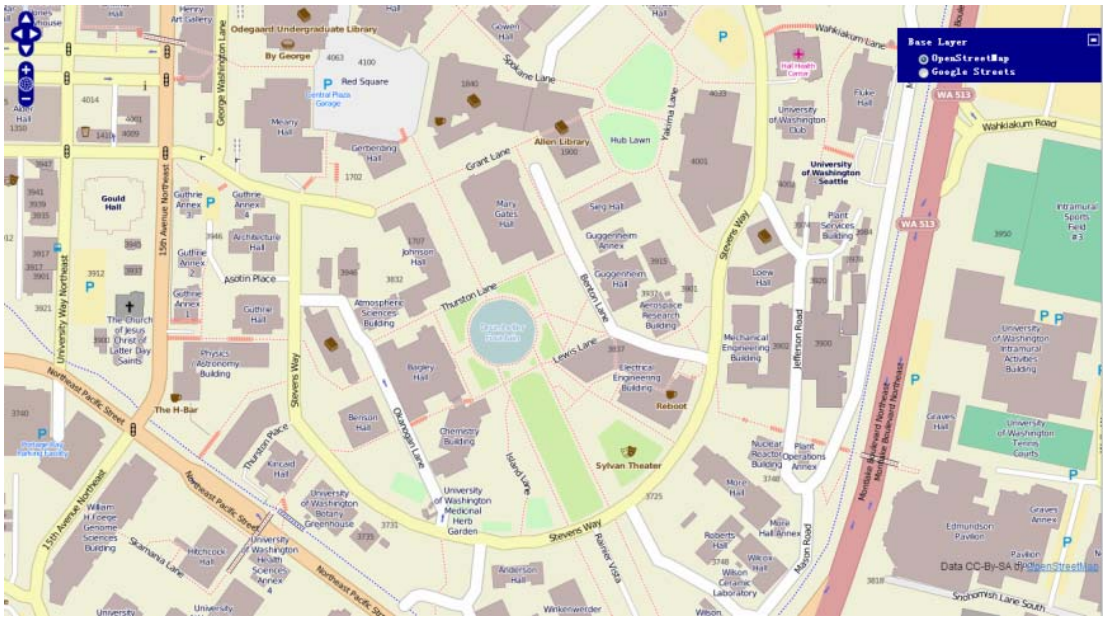


Figure 5-1 High resolution OpenStreetMap near the University of Washington

Figure 5-4 shows how clients dynamically interact with OpenStreetMap in the DRIVE Net system and the backend processes. When a Web server receives a client's 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. The Web server then passes the map contents to clients. On 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 on the client side that helps users interact with dynamic maps from disparate services. A number of 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 5-5 demonstrates.

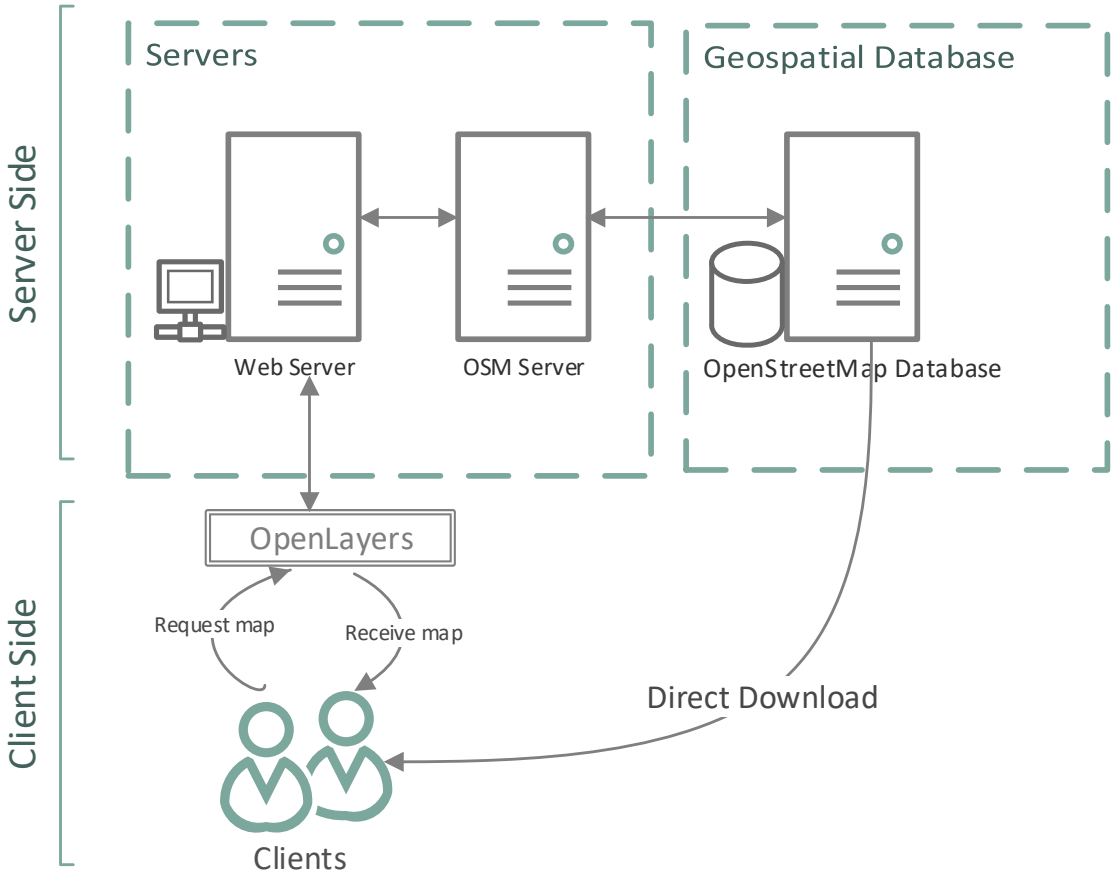


Figure 5-2 Communication mechanism for OpenStreetMap

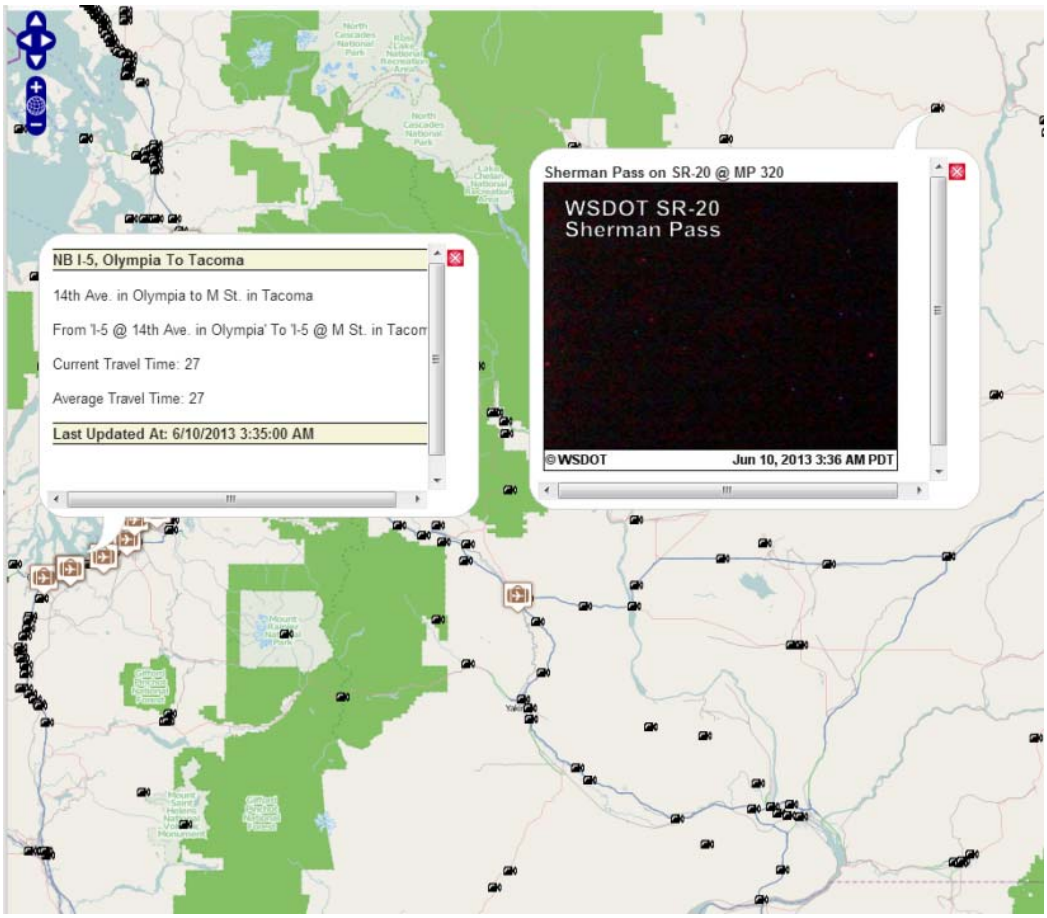


Figure 5-3 Multiple Layers on Top of a Map

5.4.2 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 Core Team, 2013). With thousands of active contributors from academia, R continues to evolve with the latest efficient and innovative algorithms. Meanwhile, R provides excellent tools for creating graphics, which enable better data visualization. Rserve, a TCP/IP server connecting to R, integrates R into the DRIVE Net 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 5-6 demonstrates. By integrating R and its countless statistical and graphic packages, DRIVE Net offers an easy and customizable

interface for performing complex analysis and data visualization for users, even those without any background knowledge of R scripts.

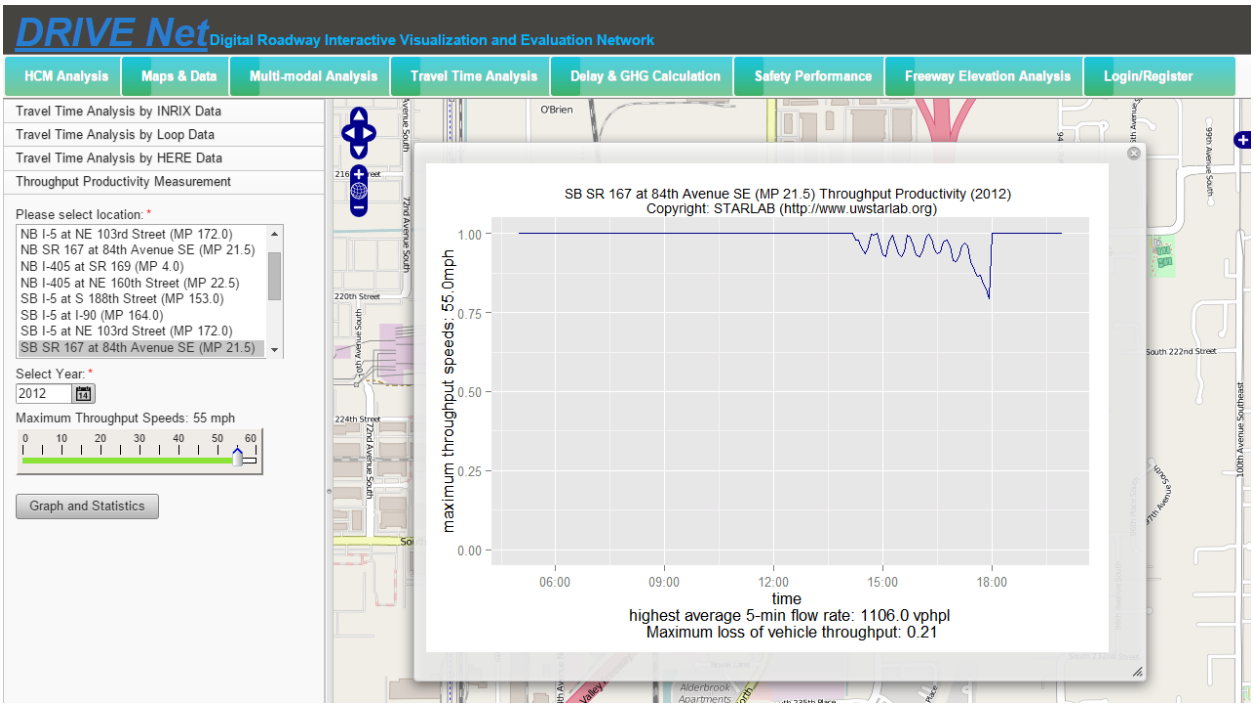


Figure 5-4 Travel time performance measurement

5.4.3 Role-Based Access Control

Role-based access control (RBAC) is an approach to restricting system access to authorized users and is also a policy-neutral access control mechanism defined around roles and privileges (Ferraiolo and Kuhn, 1992; Sandhu, et al. 1996). The components of RBAC, such as role-permissions, user-role, and role-role relationships, make it simple to assign access rights to users.

Within an organization/platform, roles are created for various functions. The permissions to perform certain operations are assigned to specific roles. Users are assigned particular roles, and through those role assignments they acquire the computer permissions to perform particular computer-system functions. Since users are not assigned permissions directly, but only acquire them through their role (or roles), management of individual user rights becomes a matter of

simply assigning appropriate roles to the user's account; this simplifies common operations, such as adding a user or changing a user's department.

Three primary rules are defined for RBAC:

(1) Role assignment: A subject can exercise a permission only if the subject has selected or been assigned a role.
(2) Role authorization: A subject's active role must be authorized for the subject. With rule (1) above, this rule ensures that users can take on only roles for which they are authorized.
(3) Permission authorization: A subject can exercise a permission only if the permission is authorized for the subject's active role. With rules 1 and 2, this rule ensures that users can exercise only permissions for which they are authorized.

Additional constraints may be applied as well, and roles can be combined in a hierarchy in which higher-level roles subsume permissions owned by sub-roles.

With the concepts of role hierarchy and constraints, one can control RBAC to create or simulate lattice-based access control (LBAC). Thus RBAC can be considered to be a superset of LBAC.

When defining an RBAC model, the following conventions are useful:

- S = Subject = A person or automated agent
- R = Role = Job function or title which defines an authority level
- P = Permissions = An approval of a mode of access to a resource
- SE = Session = A mapping involving S, R and/or P
- SA = Subject Assignment
- PA = Permission Assignment
- RH = Partially ordered Role Hierarchy. RH can also be written: ≥ (The notation: x ≥ y means that x inherits the permissions of y.)
    - A subject can have multiple roles.
    - A role can have multiple subjects.

- A role can have many permissions.
- A permission can be assigned to many roles.
- An operation can be assigned many permissions.
- A permission can be assigned to many operations.

A constraint places a restrictive rule on the potential inheritance of permissions from opposing roles; thus it can be used to achieve appropriate separation of duties. For example, the same person should not be allowed to both create a login account and to authorize the account creation.

Thus, using set theory notations,

- $PA \subseteq P \times R$ and is a many to many permission to role assignment relation.
- $SA \subseteq S \times R$ and is a many to many subject to role assignment relation.
- $RH \subseteq R \times R$

In DRIVE Net, a mechanism of RBAC is implemented for access control securely, as shown in figures 5-7 to 5-10.

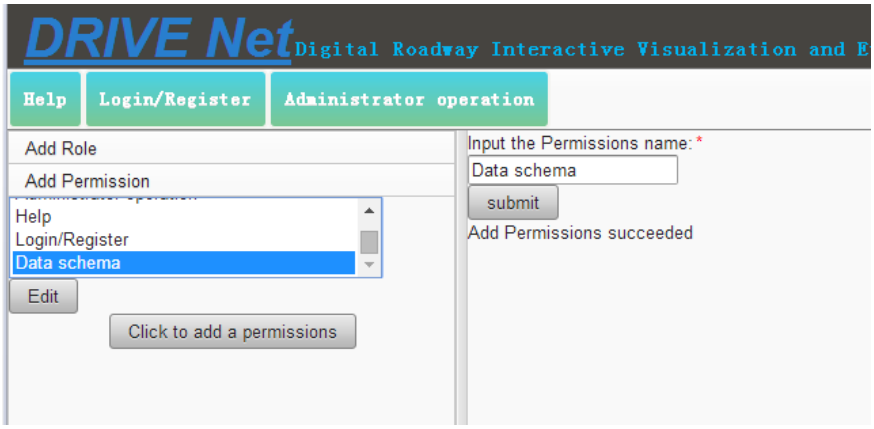


Figure 5-5 Add permission

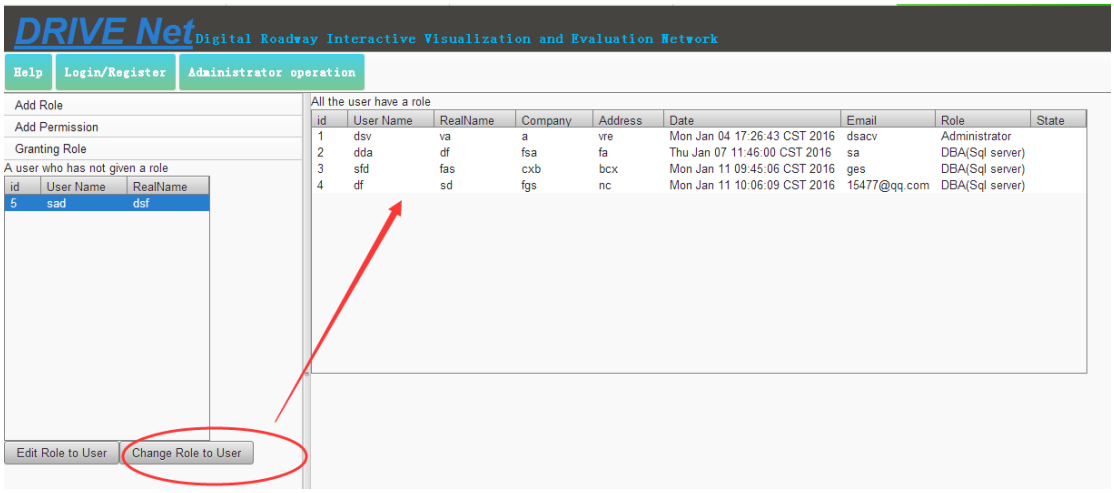


Figure 5-6 Change role

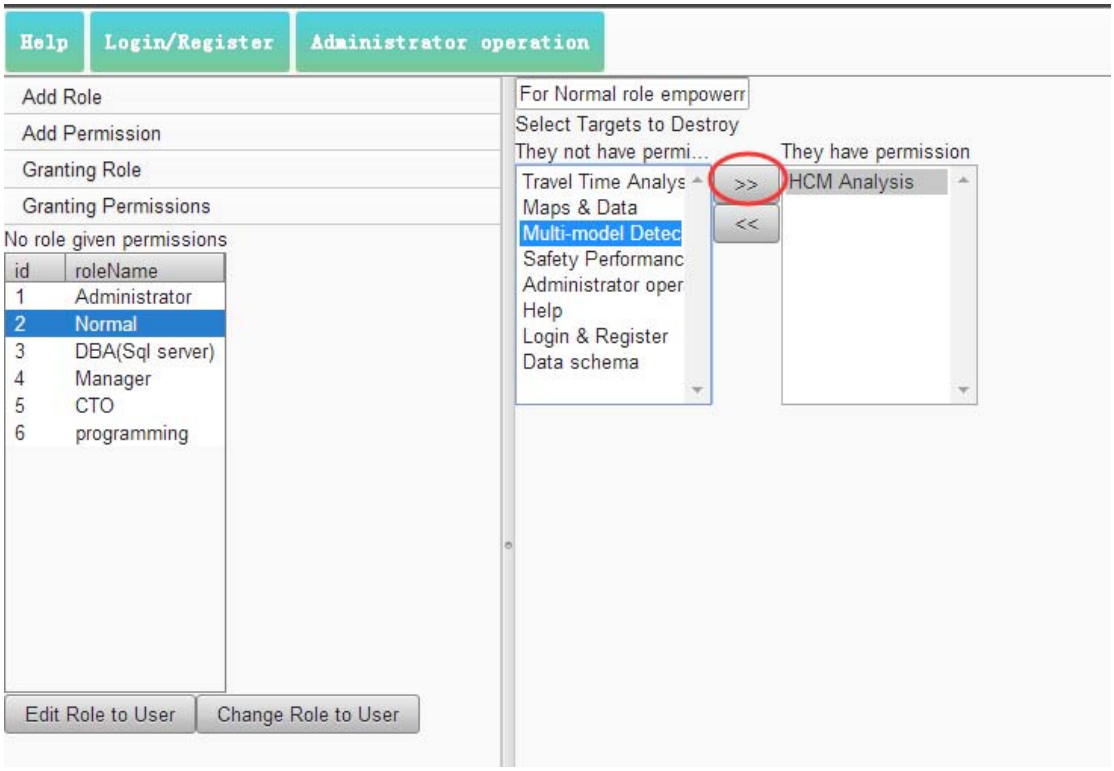


Figure 5-7 Edit/change role to user

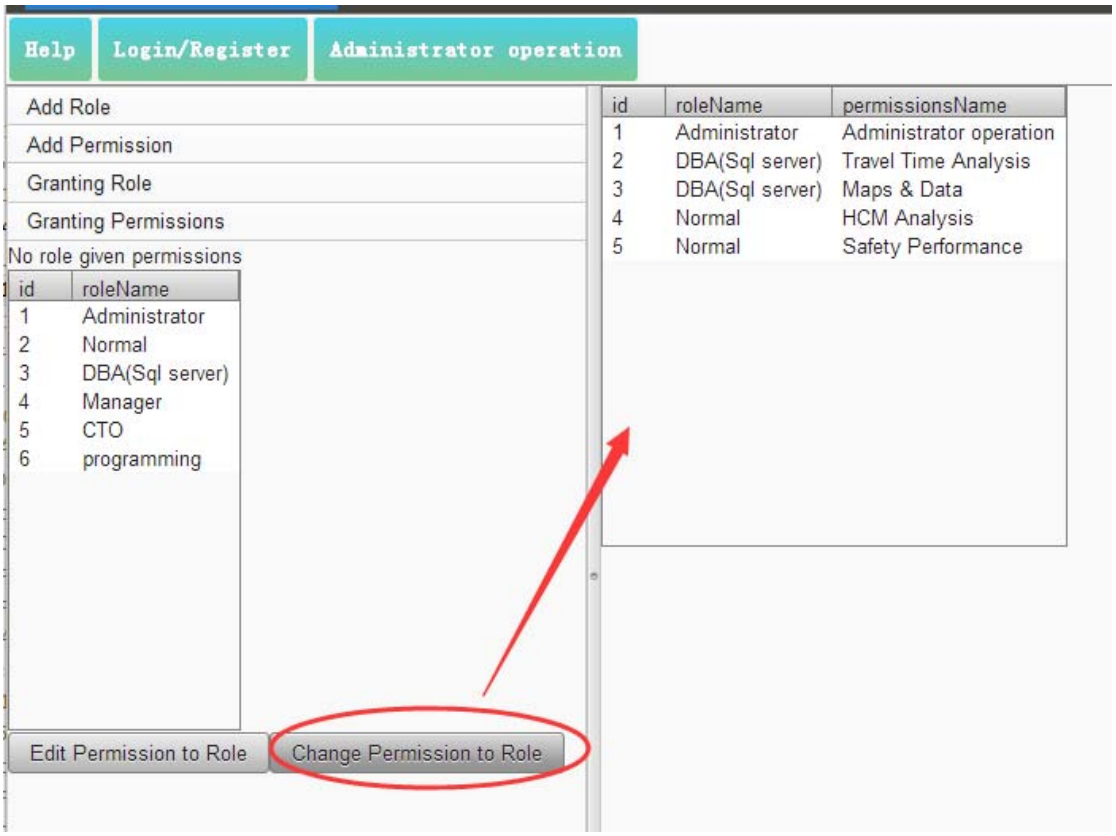


Figure 5-8 Edit/change permission to role

5.4.4 Data Source Connection Pool

A data source connection pool is an efficient way to store, organize, and retrieve data for most applications using a relational database. J2EE[1] components access relational databases through the Java Database Connectivity (JDBC) API. For information on this API, see http://java.sun.com/docs/books/tutorial/jdbc

In the JDBC API, databases are accessed via *DataSource* objects. A *DataSource* has a set of properties that identify and describe the real world data source that it represents. These properties include information such as the location of the database server, the name of the database, the network protocol used to communicate with the server, and so on. In the Application Server, a data source is called a JDBC resource.

Applications access a data source by using a connection, and a *DataSource* object can be thought of as a factory for connections to the particular data source that the *DataSource* instance represents. In a basic *DataSource* implementation, a call to the getConnection method returns a connection object that is a physical connection to the data source.

If a *DataSource* object is registered with a Java Naming and Directory Interface (JNDI) naming service, then an application can use the JNDI API to access that *DataSource* object, which can then be used to connect to the data source it represents.

DataSource objects that implement connection pooling also produce a connection to the particular data source that the *DataSource* class represents. The connection object that the getConnection method returns is a handle to a PooledConnection object, rather than a physical connection. An application uses the connection object in the same way that it uses a connection. Connection pooling has no effect on application code except that a pooled connection, like all connections, should always be explicitly closed. When an application closes a connection that is pooled, the connection is returned to a pool of reusable connections. The next time getConnection is called, a handle to one of these pooled connections will be returned if one is

[1] **J2EE** is a platform-independent, Java-centric environment from Sun for developing, building and deploying web-based enterprise applications online. The **J2EE** platform consists of a set of services, APIs, and protocols that provide the functionality for developing multitiered, web-based applications.

available. Because connection pooling avoids creating a new physical connection every time one is requested, it can help applications run significantly faster.

The Application Server is distributed with a connection pool named *DerbyPool*, which handles connections to the Derby database server.

In DRIVE Net, the application of a data source connection pool obviously reduces exceptions and errors when a relational database is accessed, which then improves reliability of the system and thereby enhances system performance to store and retrieve data.

Chapter 6 Multi-Modal Data and Visualization

6.1 Pedestrian Travel Experiment

Automatic pedestrian data collection has been challenging because of the freedom of pedestrians' movements and the lack of effective pedestrian sensors. Presently, pedestrian data collection relies largely on manual counts or video images. These approaches are both expensive and time consuming. To address this issue, the research team developed a mobile sensing approach for collecting pedestrian movement data. This approach will become increasingly attractive because of the ubiquitous use of mobile devices and their frequent need to communicate wirelessly. By capturing mobile devices' Media Access Control (MAC) addresses and re-identifying them, the movements of people carrying those devices can be identified. In this task, a mobile app was developed for use by volunteers who were willing to help collect pedestrian data. The app will turn a volunteer's mobile device into a moving sensor. The sensor will collect MAC addresses and their timestamps and then send these data, together with the volunteer's GPS location data, to the DRIVE Net server computer at the STAR Lab. These data will be processed by a computer module that implements a pedestrian trajectory reconstruction algorithm, developed in this study on the DRIVE Net platform, to estimate the routes of the detected pedestrians.

6.1.1 Problem Statement

Present pedestrian data collection approaches are limited to surveys, which are either administered on location or via broad distribution; manual counts, which involve field data collection by personnel; or automatic spot counts, achieved by either infra-red trip-line sensors or, in the case of cyclists, inductance loops. Video-based data collection methods that are capable of counts as, well as localized route choice, are also under development (Kong et al., 2006 and Malinovskiy et al., 2008). Aside from expensive, stated preference surveys, none of these approaches provides network-wide travel information. Furthermore, because of the costs of many of these approaches, communities often conduct studies only annually, picking a particular day of the year to act as a surrogate for overall performance (Alta Planning and Design, 2011).

Not only is this approach likely to produce non-representative results because of climate variations, but it also does not provide a clear trend line that can be analyzed to determine effective improvements in infrastructure or policy.

Development of a cost-effective data collection paradigm that relied on existing mobile phone infrastructure would alleviate many of these concerns and provide continuous, rich data. This is a chance to quickly address the current disconnect between community planning and available, active travel knowledge, while opening doors for additional investigations into epidemiological issues, cultural behavior, economic impacts, and community evacuation strategies.

6.1.2 Pedestrian Trajectory Reconstruction

Point sensor data are limited to providing the behavior of a given network in just a few sample points. Re-identification approaches effectively allow one to study the entire network as a whole. Some of the most available and important re-identification-based data include origin-destination pair data, which are key components in both long- and short-term forecasting efforts. These data have traditionally been collected with surveys; however, the increasing capacity to reliably re-identify individuals automatically by using the approaches described above is allowing this information to be collected without the subject's knowledge or input. This allows the collection of observed preference (instead of stated preference). However, it also relies on implied consent (at best) to collect such data. Because many of the identifiers collected are unique, it becomes relatively easy to tie a particular device to a particular point in space-time. Furthermore, since collecting origin-destination data primarily involves determination of home and work locations, it becomes increasingly easy to tie an individual to a particular device, thus violating their locational privacy.

Besides origins and destinations, imputation of intermediate points is also of interest, in particular when route choice, infrastructure effectiveness, and road pricing questions are studied. Imputation of intermediate points allows one to create trajectories, or travel diaries, for each observed entity within the network. This information has great potential for use in the new generation of activity-based models currently being built and used as transportation and land-use

forecasting tools. However, the imputation of trajectories will yield more issues related to compromising individual privacy. That is, in addition to knowing home and work locations, it is potentially possible to impute places of worship, shopping habits, and a host of other individual behavior characteristics. Because many models rely on a variety of indicators to improve predictive power, there is a greater conflict between building accurate models and imputing or otherwise obtaining increasingly invasive data. As the possibility of MAC-based, network-wide re-identification becomes more apparent, these privacy issues must be addressed. In addition, the inherent uncertainties within the data collection method must be mitigated. A framework for pedestrian trajectory reconstruction was developed as an important module in the DRIVE Net system.

6.2 Public Transit Data

Public transit is a convenient, safe, and comfortable means of transportation. In DRIVE Net, we designed a public transit data visualization panel for users. The data for this panel are obtained from the OneBusAway API, currently providing real-time bus information in the Puget Sound region. Two searching functions were implemented in this panel, which are bus route searching and bus stop searching. Users can input the number of a bus route to search the information on the transit routes or stops. The application step is shown as follows.

(1) Input a route number.

Under the Multi-Modal Analysis tab, select the Public Transit Data, then input the route number in the text field (Figure 6-1).

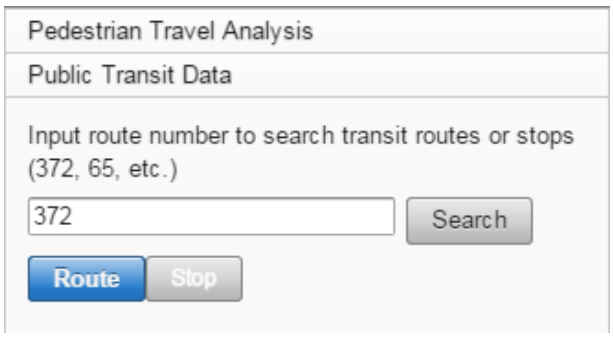
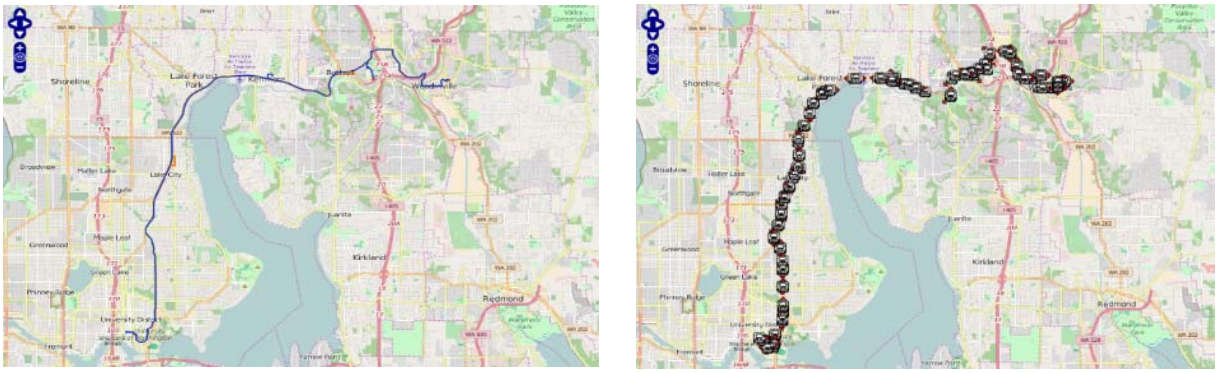


Figure 0-1 Input route number under public transit data

(2) Display the detailed information.

For example, the user inputs the route number 372. If the “Route” button is selected, the route will be shown on the map. If the “ Stop” button is selected, all the bus stops will be shown on the map (Figure 6-2).



(a) **(b)**

Figure 0-2 Information about the route and the stops

In the design, each bus stop icon has a click listener added. Thus, if any bus stop icon is clicked, the detailed static information will be displayed in a pop-up window, such as the stop name and stop ID as well as dynamic information including the predicted arrival time of the next bus (Figure 6-3).

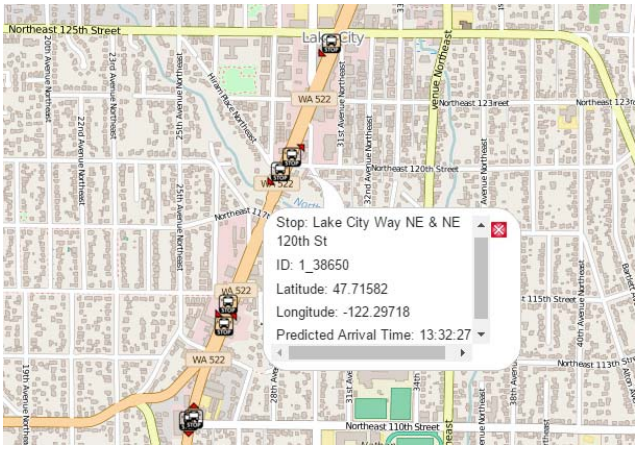


Figure 0-3 Detailed information about the stops

6.3 Park and Ride data

Park and ride data can be obtained from DRIVE Net. Currently, the park and ride data on DRIVE Net are static data for visualization and basic information searching. By searching by county name, city name, number of spaces, or all the park and ride locations, users can get the park and ride locations they are interested in shown on the map.

(1) Input the information

After click a tab (by County, by City, by # of Spaces, All), input the name of the county (King, Yakima, Spokane, etc), name of the city (Seattle, Arlington, Vancouver, etc.) or approximate number of spaces (10, 201, etc.) (Figure 6-4).

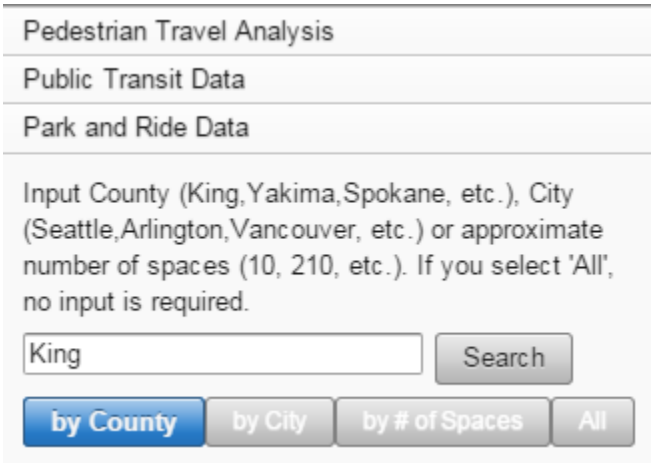
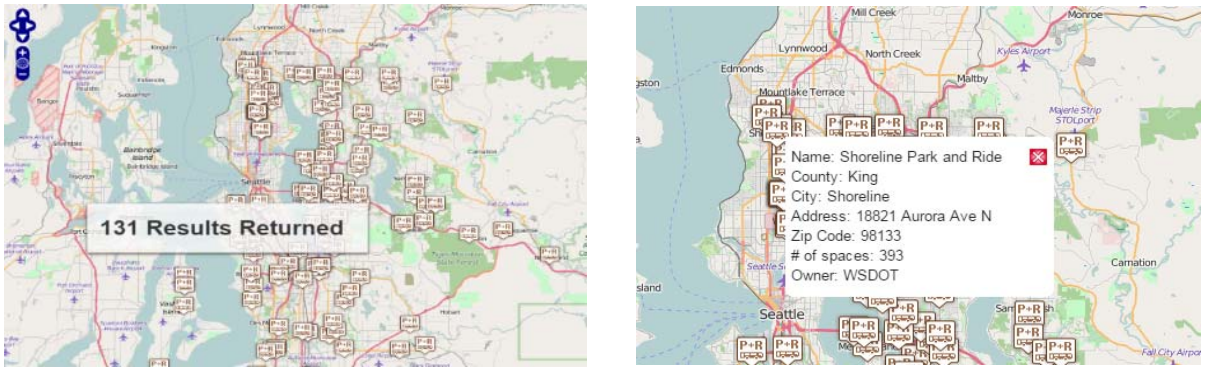


Figure 0-4 Input the county, city, or number of spaces under park and ride data

(2) Display the detailed information

If King is input, then the button of by County is selected, and the available park and rides will be shown on the map. If a user clicks a particular park and ride, then its name, county, city, address, zip code, number of spaces and owner will be shown (Figure 6-5).



(a) **(b)**

Figure 0-5 Information about available park and detailed information

6.4 Car2go Data

Unlike the traditional car rental system, Car2Go does not necessarily need to be rented from or returned to a certain car rental place. Instead, the rental and return can be done at any place, which is very convenient and flexible. Through the official Car2Go phone app (Figure 6-6), users can view the available Car2Go vehicles' locations and reserve the nearest vehicle.

Figure 0-6 Installing the App in a smartphone enables Car2Go service

In DRIVE Net, users can select the region, such as Seattle, Austin, or Denver, where available cars will be shown on the map. By clicking a car icon, its name, engine type, address and fuel level will be shown in a pop-up window (see Figure 6-7).

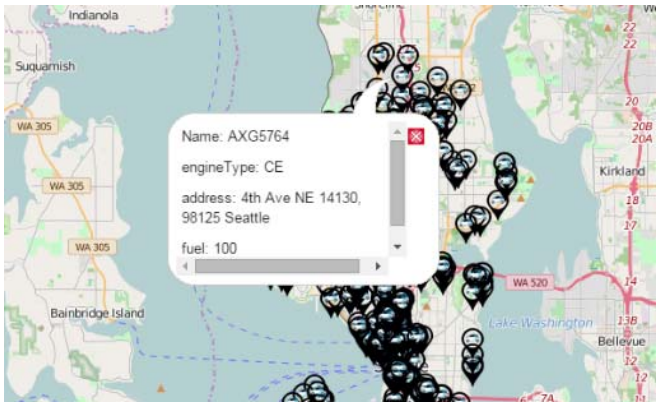


Figure 0-7 The pop-up window shows the detailed information of a selected available car

6.5 Bicycle and Pedestrian Count Data

Bicycle and pedestrian count data are available on DRIVE Net under the "Bicycle & Pedestrian Count Data" panel. The bicycle and pedestrian count data in a specified location (usually on a trail) can be downloaded to a local computer for future research and application. By selecting a location name in the dropdown menu, then selecting the start and end date, the bicycle and pedestrian count data will be downloaded in a Excel file. Figure 6-8 shows the interface.

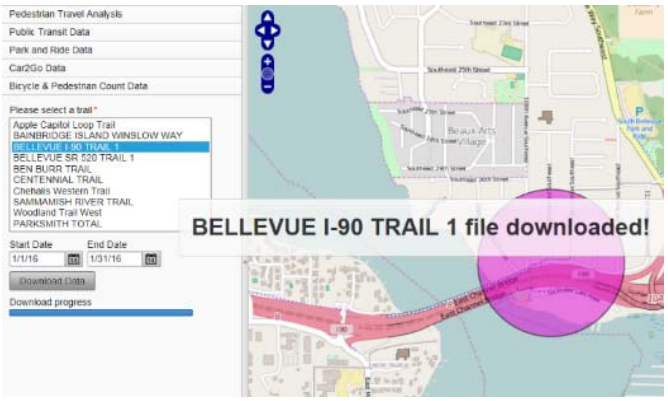


Figure 0-8 Interface of the bicycle and pedestrian count data download

6.6 Ferry Data

Ferries are an important part of some traffic systems, especially in Washington state. In the DRIVE Net system, ferry terminal and real-time ferry data are available under the panel named

WSDOT Real-Time Data. In this panel, if a user clicks the checkbox for Ferry, all the ferry terminals and ferries belonging to WSDOT will be displayed on the map. Blue icons represent ferries, and the icons with white background represent ferry terminals. If the user clicks a ferry icon, a pop-up window showing the ferry's real-time information (e.g., location, speed, and available space) will be displayed (see Figure 6-9).

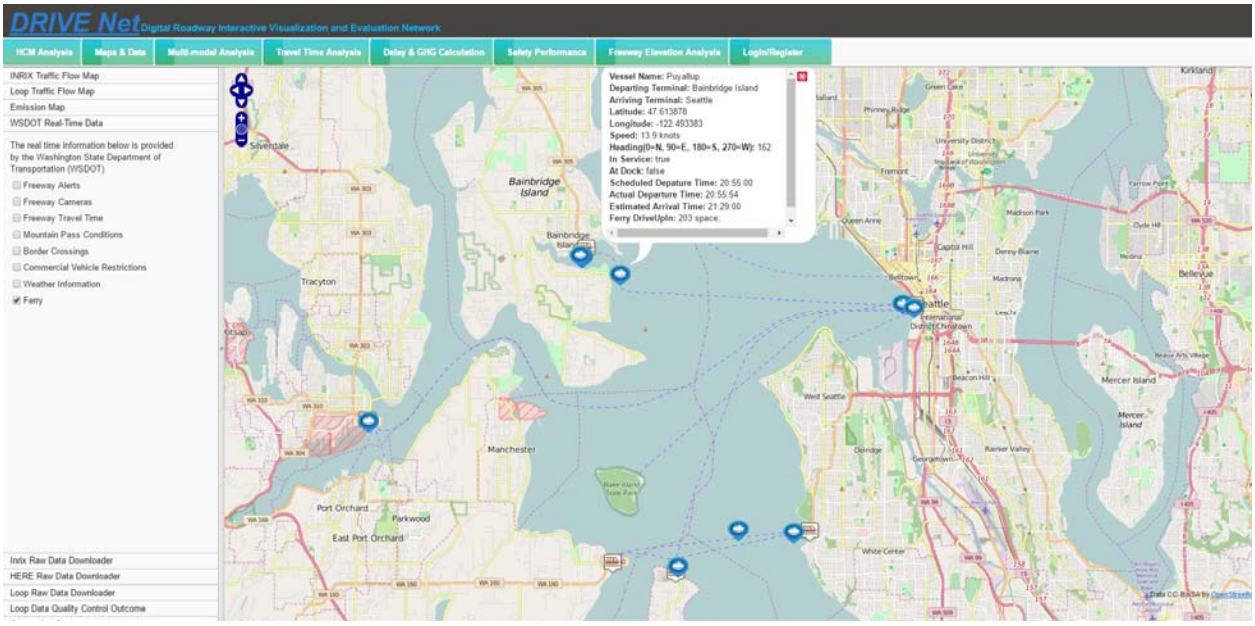


Figure 0-9 Ferry real-time information

Chapter 7 Travel Time Analysis

7. 1 Introduction

Efficient traffic control and successful traffic management require the availability of precise data on current traffic conditions. Besides detector data collected at various points in the street network, measured travel times are playing an increasing role because they are more accurate indicators of traffic disturbances. Travel times reflect all disruptive factors on a given route, even those incidents or conditions that cannot be detected by an individual detector (Siemens, 2016). Therefore, WSDOT uses three different data sources—loops, HERE and INRIX—to measure travel time for traffic management.

With the processed detector data sets, statistics for WSDOT's Gray Notebook (GNB) can be calculated by leveraging the power of eScience. This section documents the development of GNB statistics from the DRIVE Net system. A major component of the GNB are the freeway performance monitoring results that WSDOT annually collects statewide. To meet its goals, WSDOT has purchased private sector, probe-based speed data to assist in generating the WSDOT congestion report. WSDOT also uses loop detector data from 6800 loop detectors, gathered from 26 commuter routes in the Puget Sound area, to calculate congestion conditions. Travel time analysis and throughput productivity evaluation are two important sections in the WSDOT congestion report.

For commuters, travel times and travel reliability are two important performance indicators. Key information includes the average peak travel time, the 95 percent reliable travel time, the duration of congestion, the percentage of weekdays when average travel speeds are below 36 mph, and the maximum throughput travel time index (MT^3I). The congestion performance of each route for the current year is compared with that of a baseline year. The average peak travel time is the average travel time during the peak 5-minute intervals for all weekdays of a whole year. The duration of congestion is defined as "the period of time during which average trip speeds fall below 45 mph (75 percent of the posted speed)" (WSDOT, 2013). MT^3I is used to compare travel times on routes with different lengths, and it can be calculated as the ratio between average peak travel time and maximum throughput speed travel time.

Maximum throughput speed travel time can be obtained by dividing the length of a route by the maximum throughput speed. However, in reality, the maximum throughput speed is dynamic and hard to acquire because of multiple factors. To simplify the calculation of freeway congestion metrics, 85 percent of the posted speed is adopted as the maximum throughput speed.

For the throughput productivity evaluation, vehicle throughput is the total vehicle hourly volume on a segment at a point location, and the lost throughput productivity is thus defined as "the difference between the highest average 5-minute flow rate observed during the year and the flow rate that occurs when vehicles travel below the maximum throughput speeds" (WSDOT, 2013).

DRIVE Net phase 2 adds several new functions to satisfy WSDOT's requirements. There are three options on the GNB Calculation screen: travel time analysis using INRIX data, travel time analysis using loop data, and throughput productivity measurements, which are calculated on the basis of the INRIX data and loop data. In DRIVE Net phase 2, HERE data have been added to the system to enhance this platform's travel time analysis performance, and a new user-defined corridor selection function can help users choose corridors more conveniently.

7.2 WSDOT Gray Notebook Statistics Implementation on DRIVE Net

The traffic information collected from loop detectors is the main source of data for computing travel time on corridors, as well as vehicle throughput productivity. The critical steps for estimating travel time are summarized as follows:

Step 1 – Corridor Segmentation

In Step 1, corridors are segmented on the basis of cabinet locations. The midpoints of the cabinets are used to naturally break the corridor down into segments. For instance, as Figure 7-1 depicts, the corridor is divided into three segments, d_1, d_2, and d_3, by splitting it up at the midpoints of three cabinets (Figure 7-1). The speed of each segment is then taken from the nearest loop detectors.

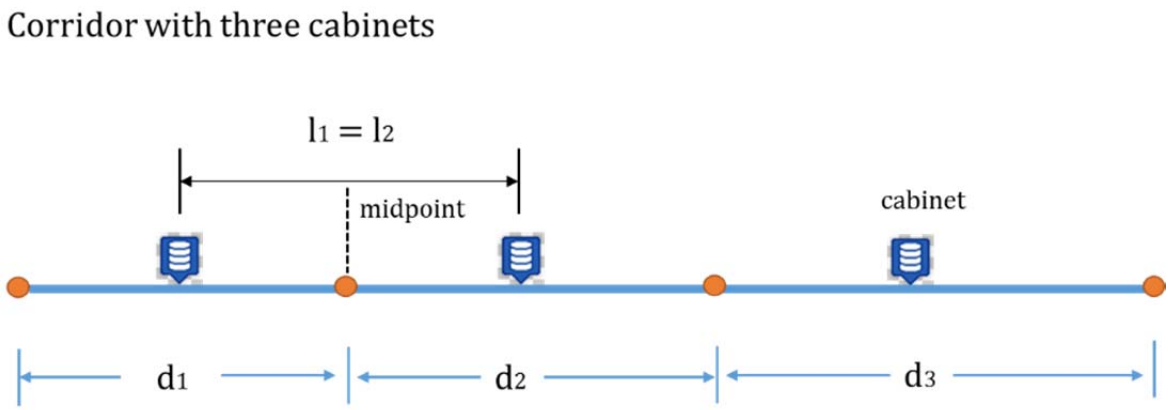


Figure 0-1 Corridor segmentation

Step 2 – Speed Calculation

To calculate the travel time, first the travel time on each selected corridor must be calculated. Since there are three different data sources, the travel time is calculated on the basis of three methodologies.

Because loop data contain only loop scanning and volume information, the WSDOT speed calculation formula is used (WSDOT, 2016):

$$Loop\ Speed = \frac{Volume}{Occupancy \times T \times g} \tag{7-1}$$

where T, which is the length of interval in hours, equals $\frac{20}{3600}$, Occupancy equals $\frac{ScanCount}{1200} \times 100$, and g, which is the speed factor, equals 2.4.

For INRIX data, the original data source has the speed information. Therefore, the travel time is calculate from the formula:

$$Travel\ time = \frac{Distance}{Speed} \tag{7-2}$$

For HERE data, the original data source has the travel time information. Therefore, no additional calculations for travel time are necessary.

Step 3 – Five-Minute Interval Travel Time Computation

The traffic data are aggregated into speed values in 5-minute intervals. The lengths of segments are computed on the basis of the mileposts of the cabinets. Once the speed and length for segments are known, the travel time on entire corridor can be estimated by summing all the segments' times. The system further prepares the 5-minute travel time of the corridors for all weekdays in the year selected.

$$corridor\ travel\ time = \sum_{i=1}^{n} d_i/s_i \quad (7\text{-}3)$$

where

$$n: total\ number\ of\ segments\ on\ the\ corridor$$

$$d_i: distance\ of\ segment\ i$$

$$s_i: speed\ of\ segment\ i$$

Step 4 – Determination of Peak Time

For each 5-minute interval between 5:00 to 10:00 (morning) or 14:00 to 20:00 (evening), the system averages travel time for all weekdays of the whole year. The 5-minute time slot with the highest observed average travel times for morning or evening is then determined as the peak time of the commuter AM or PM rush, respectively.

Step 5 – Travel Time Reliability Analysis

Once the peak 5-minute interval has been determined, average travel time, 50th percentile travel time, 80th percentile travel time, 90th percentile travel time, and 95th percentile travel time can be found from the data set prepared in Step 2. The system further calculates the MT^3 index, peak period VMT, and duration of congestion to compare travel time among corridors with different lengths.

Similarly, for throughput productivity analysis, the cabinets close to the 16 monitored locations are used to provide volume and speed information. For each location, the system averages the 5-minute flow rate as well as speeds for all weekdays in the year. The highest observed average 5-minute flow rate, V_o, passing through a location is then defined as the

optimal throughput. Using this value as the basis, throughput productivity is computed with Equation (7-4).

$$throughput\ productivity = \begin{cases} 1, & speed \geq maximum\ throughput\ speed \\ 1 - V/V_o, & speed < maximum\ throughput\ speed \end{cases} \quad (7\text{-}4)$$

where

$$V_o: optimal\ throughput$$

$$V: 5 - minute\ flow\ rate$$

7.3 Travel Time Analysis Metrics

Commute trip analysis refers to WSDOT's calculation of various congestion performance measures. WSDOT transforms traffic data into performance measures that tell the commute congestion story for urban travel. In the DRIVE Net system, the following metrics are used to analyze travel time and commute trips as WSDOT does, according to the Corridor Capacity Report (WSDOT, 2016).

- **Daily commutes**
  - Peak period
  - Peak 5 minutes of commuter rush
- **Travel times**
  - Average peak travel time
  - Travel time at maximum throughput speed
  - Maximum Throughput Travel Time Index (MT_3I)
  - Travel time index
  - Planning time index
  - Buffer index
- **Reliability**
  - Average travel time
  - Median travel time (50th percentile)
  - 80th percentile reliable travel time
  - 90th percentile reliable travel time
  - 95th percentile reliable travel time
- **Congestion**
  - Duration of congestion
  - Commute congestion cost

The DRIVE Net system, based on the requirements and previous work in DRIVE Net phase 1, calculates the majority of the metrics listed above, which are the most important metrics for WSDOT's corridor capacity analysis, shown in Figure 7-2. Below are descriptions of how those metrics are calculated in the travel time analysis function of the DRIVE Net system.

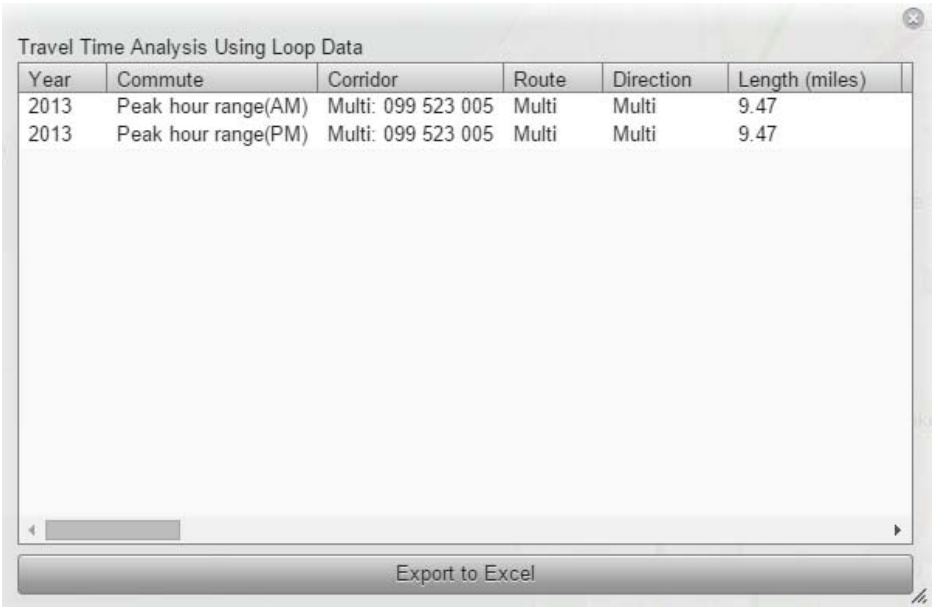


Figure 0-2 User defined travel time analysis display table

- **Average Peak Travel Time**

Average peak travel time is the longest travel time within the morning and evening peak periods. The corresponding 5-minute intervals become the peak 5-minute intervals as defined above.

$$Average\ peak\ travel\ time = \frac{Trip\ length}{Average\ speed\ _{5-min}} \quad (7\text{-}5)$$

- **Average Travel Time**

Average travel time (the mean) is the average of all the recorded travel times. This measure describes the "average" experience on the road that year.

- **50th Percentile Travel Time**

50th percentile travel time (the median) is the middle value of all the recorded travel times. The median is not affected by very long travel times as an average is, so it gives a better sense of actual conditions.

- **80th Percentile Travel Time**

80th percentile travel time describes conditions under which the traveler is on time four out of five weekday trips. WSDOT uses this percentile to track changes in reliable travel times over the years at a finer level, to better evaluate operational improvements.

- **90th Percentile Travel Time**

90th percentile travel time means that 90 percent of all the recorded travel times are shorter than this duration.

- **95th Percentile Travel Time**

95th percentile travel time means that the traveler will be on time approximately 19 out of 20 weekday trips. WSDOT uses this percentile as its key reliability metric.

- **Maximum Throughput Travel Time Index (MT_3I)**

Maximum Throughput Travel Time Index (MT3I) helps compare travel times on routes of different lengths. The MT3I incorporates the expected travel time under maximum throughput conditions and the travel time at the peak 5-minute interval, thereby taking into account the length of the route. An MT3I of 1.0 indicates a highway operating at maximum efficiency. As the MT3I value increases, travel time performance deteriorates.

$$Maximum\ throughput\ travel\ time\ index(MT^3I) = \frac{Average\ travel\ time\ _{Peak\ 5-min}}{Travel\ time\ _{Max.\ throughput\ speed}} \qquad (7\text{-}6)$$

- **Duration of Congestion**

Duration of congestion (DOC) captures the amount of time that a commute corridor typically experiences speeds slower than 75 percent of the posted speed (45 mph when the posted speed is 60 mph) on an average weekday for the analysis period.

$$Duration\ of\ congesiton\ \ = \sum(Time\ for\ all\ intervals\ with\ speeds < 45\ mph) \qquad (7\text{-}7)$$

- **Commute Congestion Cost**

Commute congestion cost is the economic impact of time and fuel wasted from extra travel time incurred by drivers during congested periods. WSDOT calculates commute congestion cost by applying monetary values to the extra travel time and vehicle operating costs drivers experience during congested periods.

$$Commute\ congestion\ cost = \sum \left(Average\ travel\ time_{\ 5\ min} - Travel\ time_{\ threshold\ speed}\right) \times Traffic\ volume_{\ 5\ min} \times Cost\ per\ minute \quad (7\text{-}8)$$

Commute congestion cost is computed for every 5-minute interval within the time that a particular commute is experiencing congestion. The commute congestion cost computation is based on the duration of congestion calculation for a particular commute route. The default factor of cost per minute is set as \$21.9/hour, which is converted to be \$0.365/minute/vehicle. This factor's value can be changed in the user defined loop data travel time analysis interface of the DRIVE Net system, shown in Figure 7-3.

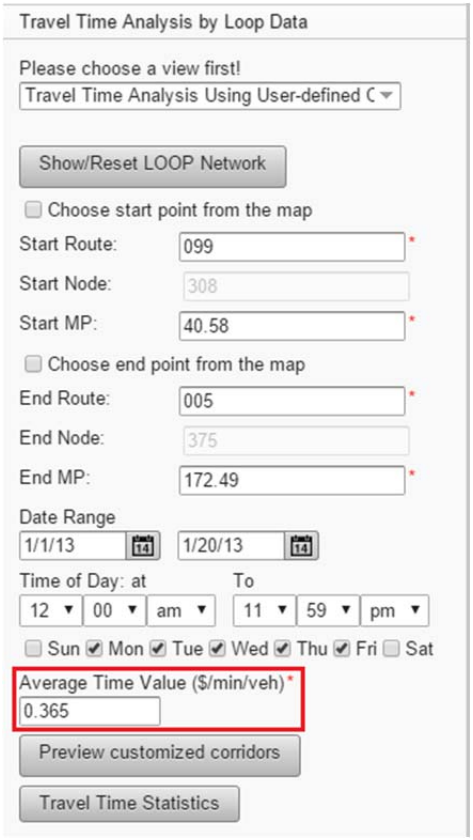


Figure 0-3 User defined loop data travel time analysis interface

7.4 Travel Time Analysis by Loop Data

7.4.1 GNB Trips

DRIVE Net can be used to estimate travel time and reliability measures for 26 Puget Sound commuter routes. In the loop data travel time analysis section, each of these routes is separated into two categories: GP lanes and HOV lane. In INRIX and HERE data travel time analysis section, the functions are similar. In addition, DRIVE Net can be used to estimate the average travel speed for a corridor, which is used to estimate the number of days for which congestion is present at any given time of day. For travel time using INRIX, loop, or HERE data, users start by selecting a corridor and travel year (Figure 7-4, left). The stamp graph threshold selector sets the traffic speed threshold below which traffic conditions are considered to be congested. For example, on a corridor with a 60 mph speed limit, traffic may be assumed to be congested if the

speed drops below 36 mph. The scroll button will display the current speed threshold. When clicking on the Travel Time Statistics button, the system will download a summary of travel time for the selected corridor, shown in Figure 7-5.

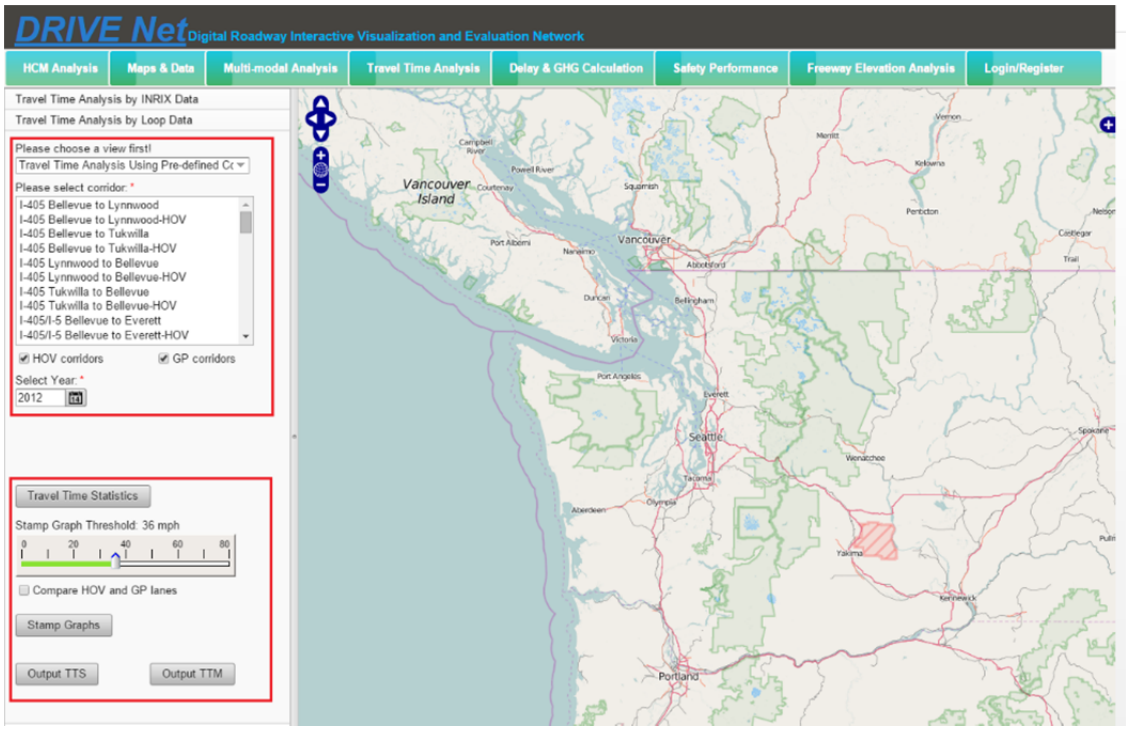


Figure 0-4 Loop travel time analysis using pre-defined corridor

				Travel times on the route at		Average travel time at peak of rush			Maximum throughput travel time		VMT during peak period	Duration of congestion (how long the average speed is below 45mph)			95th percentile		
Route	Direction of Travel	Length of route	Peak time of commuter AM rush	Posted speed	Maximum throughput speed	2010	2012	%Δ	MT^3 Index 2010	2012	%Δ in VMT	2010	2012	Δ	2010	2012	%Δ
To Seattle																	
I-5 Everett to Seattle-HOV	SB																
I-5 Federal Way to Seattle-HOV	NB																
I-90/I-5 Issaquah to Seattle-HOV	WB/NB																
SR 520/I-5 Redmond to Seattle-HOV	WB/SB																
I-5 SeaTac to Seattle-HOV	NB																
I-405/I-90/I-5 Bellevue to Seattle-HOV	SB/WB/NB																
I-405/SR 520/I-5 Bellevue to Seattle-HOV	NB/WB/SB																
To Bellevue																	
I-5/I-405 Everett to Bellevue-HOV	SB																
I-405 Lynnwood to Bellevue-HOV	SB																
I-405 Tukwilla to Bellevue-HOV	NB																
I-5/I-90/I-405 Seattle to Bellevue-HOV	SB/EB/NB																
I-5/SR 520/I-405 Seattle to Bellevue-HOV	NB/EB/SB																
I-90/I-405 Issaquah to Bellevue-HOV	WB/NB	9	05:10	9	11		11			0.95			00:05			14	
SR 520/I-405 Redmond to Bellevue-HOV	WB/SB																

Figure 0-5 Travel time statistics of loop data

The DRIVE Net system provides a function to compare traffic conditions between GP lanes and HOV lanes that can be accessed by clicking on the Stamp Graph button. When the Stamp Graph button is clicked, the system displays a plot with the percentage of days with an average speed below the stamp graph threshold on the y-axis and time of day on the x-axis, as shown in Figure 7-6. This plot can also show the curve for the HOV lane for comparison.

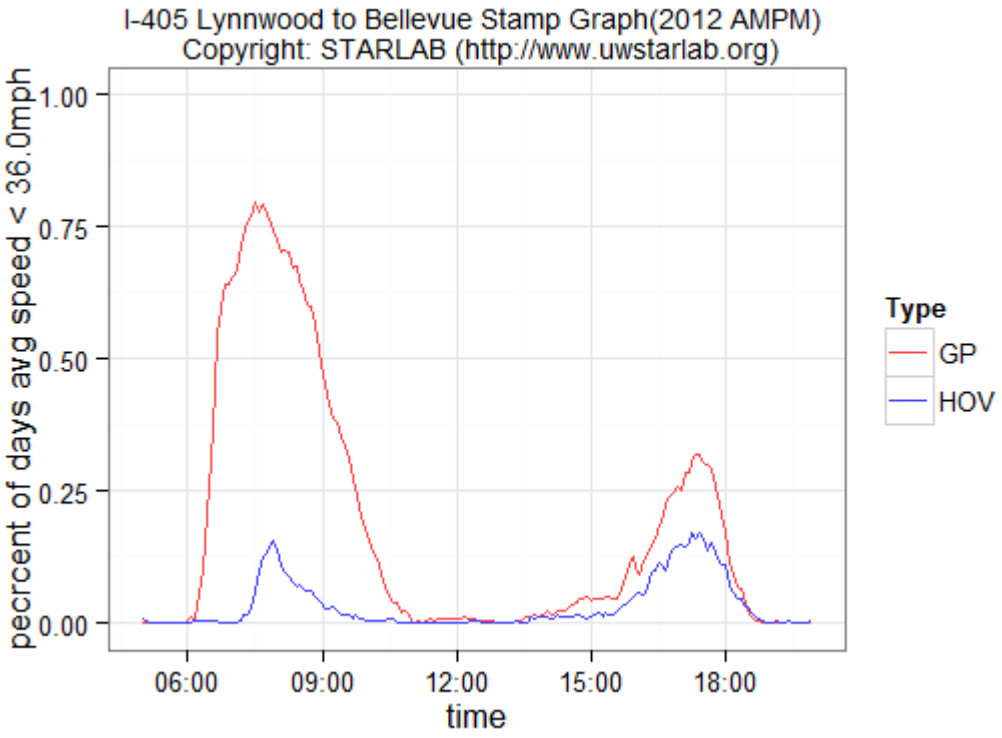


Figure 0-6 Comparison between GP and HOV lanes

Users can also generate a Travel Time Statistics (TTS) file by clicking the Output TTS button. The TTS file contains the travel time estimated for each 5-minute interval for each day selected along with travel time summaries by 5-minute interval, which is similar to the Washington State Transportation Center's (TRAC) standard. A TTS file for I-90/I-405, Issaquah to Bellevue-HOV, generated by the system is shown in Figure 7-7. The x-axis describes the summaries of travel time by 5-minute interval and y-axis is the temporal information from 5:00am to 8:00pm with 5-minute intervals.

Time	Avg TTS	Avg Speed	Min Speed	Max Speed	% cong(45.0)	% cong(36.0)	95%ile	90%ile	85%ile	80%ile
05:00	9.7836308	54.090349	29.590411	70.412034	0.003831418	0.003831418	11.285841	10.892175	10.72605	10.546754
05:05	9.5583716	55.365079	29.549693	70.750298	0.007662835	0.003831418	10.784048	10.476569	10.25753	10.031841
05:10	11.653984	45.409363	28.864682	63.406266	0.103448276	0.011494253	15.487102	14.079492	13.51062	12.837232
05:15	10.382906	50.96839	18.373303	72.378415	0.034482759	0.01532567	13.090725	12.133039	11.48262	11.243323
05:20	8.9992253	58.805062	24.998125	74.394425	0.01532567	0.007662835	10.55925	9.9489137	9.703899	9.5526791
05:25	9.3496369	56.601128	24.312392	71.09085	0.007662835	0.007662835	11.075021	10.496844	10.14109	9.944278
05:30	11.459765	46.178959	21.383318	62.682888	0.068965517	0.01532567	14.519479	13.51758	13.00223	12.512737
05:35	9.9501111	53.185336	21.244944	65.655913	0.019157088	0.011494253	11.571358	10.969144	10.65592	10.402942
05:40	10.56592	50.085556	22.241825	61.846436	0.01532567	0.003831418	12.601283	11.896278	11.44944	11.18134
05:45	9.4329199	56.101399	19.847911	68.379293	0.01532567	0.011494253	10.896924	10.27158	9.996516	9.76651
05:50	8.5802647	61.676419	29.485903	80.464071	0.011494253	0.003831418	9.7758053	9.1115297	8.901773	8.8257241
05:55	10.183877	51.964493	28.654076	62.591299	0.01532567	0.003831418	11.978628	11.326764	10.91951	10.763915
06:00	10.830493	48.862041	19.841432	63.307014	0.030651341	0.019157088	12.748805	11.935825	11.68017	11.463374
06:05	9.1814116	57.638196	25.32161	68.855768	0.019157088	0.007662835	10.280165	9.7320977	9.599095	9.4791666
06:10	8.9239449	59.301128	22.190618	67.948236	0.019157088	0.003831418	9.9821403	9.4460873	9.254427	9.1693072
06:15	9.8410817	53.774577	22.599258	63.452122	0.011494253	0.003831418	10.785598	10.397739	10.22282	10.114494
06:20	9.4105893	56.234523	33.435568	62.644032	0.011494253	0	10.3528	9.9477953	9.793928	9.6955655
06:25	8.9759269	58.9577	34.21553	72.445303	0.011494253	0	9.7924638	9.3581572	9.280949	9.1691821
06:30	9.2612626	57.141237	25.119993	65.665755	0.011494253	0.003831418	10.168216	9.7270658	9.546812	9.451247
06:35	9.0185684	58.678936	30.01047	66.461007	0.01532567	0.003831418	9.7278485	9.4280544	9.313197	9.1947857
06:40	9.2633058	57.128633	30.230965	63.482696	0.007662835	0.003831418	10.241552	9.6455104	9.518651	9.4147407
06:45	9.0887341	58.22593	25.738543	65.467811	0.007662835	0.007662835	10.056034	9.6778791	9.482124	9.3281499
06:50	9.2537051	57.187904	23.110742	63.768893	0.011494253	0.011494253	9.9287824	9.6754649	9.51245	9.4390583
06:55	9.2656158	57.114391	23.98781	62.618502	0.003831418	0.003831418	10.28957	9.7614699	9.562634	9.4447663
07:00	9.3265855	56.741023	29.979082	67.084196	0.011494253	0.003831418	10.601993	9.8683796	9.648748	9.5450538
07:05	9.2582748	57.159677	25.109805	63.536525	0.011494253	0.003831418	10.821374	10.116081	9.662524	9.5621495
07:10	9.5137206	55.624926	25.071611	74.740935	0.011494253	0.003831418	10.787939	10.317967	9.989743	9.7726142
07:15	9.74282	54.316922	20.108283	61.842544	0.011494253	0.011494253	11.463147	10.732647	10.31422	10.14922
07:20	9.7587531	54.22824	26.862149	61.662534	0.011494253	0.003831418	11.374183	10.751903	10.38957	10.158125
07:25	9.7548024	54.250202	25.74978	61.536466	0.01532567	0.011494253	11.227455	10.700305	10.34783	10.144352

Figure 0-7 Travel Time Statistics (TTS) file of loop data

The DRIVE Net phase 2 system now also can generate the Travel Time Matrix (TTM), which is also similar to TRAC's standard. A TTM file for I-90/I-405, Issaquah to Bellevue-HOV, generated by the system is shown in Figure 7-8. The x-axis describes the dates selected by users and the y-axis is the temporal information from 5:00am to 8:00pm with 5-minute intervals.

Time	2012-01-02	2012-01-03	2012-01-04	2012-01-05	2012-01-06	2012-01-09	2012-01-10	2012-01-11	2012-01-12	2012-01-13
05:00	11.0479172	9.3369654	8.98622115	9.5499179	9.9709463	9.9654674	9.2681005	9.2358436	10.281007	9.2195917
05:05	10.6985278	8.3246318	10.0201109	10.937091	10.002114	9.8369259	9.1567219	8.8442499	10.256947	10.306058
05:10	10.0102166	10.066336	11.2320932	10.883381	11.047936	12.056618	15.694365	11.960011	11.695278	14.616463
05:15	9.08089732	9.5443295	8.72729138	11.417642	8.8654666	9.2493986	10.926312	10.014113	10.11764	10.054456
05:20	9.12334616	8.6475431	8.00295047	8.8604995	9.4585036	8.5060575	8.6309554	9.1632809	8.8150539	9.7263777
05:25	10.6924032	8.4337632	10.362907	8.8091268	11.231989	9.5575921	10.401055	9.8652977	9.6481735	9.7339689
05:30	9.42558774	11.28628	11.2249471	10.45602	13.957753	10.672506	11.052627	11.482028	12.249094	10.693883
05:35	9.24214845	10.862811	10.0181509	10.063597	9.6456494	9.3721257	9.4827975	10.056053	9.2928036	10.453631
05:40	9.66023258	9.7139499	10.1163285	11.102568	10.575149	9.6198372	10.185475	9.3516139	10.302913	10.956761
05:45	8.79088837	11.025661	8.84307578	8.6145773	9.3078824	8.3865238	9.4637273	10.273747	8.8804572	8.9640918
05:50	8.74482538	8.2871535	8.19463435	8.6299018	8.5356614	8.9147255	8.4661232	8.3745082	8.4598858	8.9811549
05:55	10.5830164	9.5889476	10.1338801	10.107166	9.4844363	9.6351468	9.4809048	9.5635977	9.5374278	8.9623956
06:00	9.95633078	11.887847	10.9422637	12.755533	10.701665	11.24209	10.612512	11.123099	10.622284	11.228445
06:05	10.691317	9.3569971	9.08593206	9.4317909	9.2721517	9.7307978	9.9160062	9.1486628	8.8819839	9.388151
06:10	10.0086119	8.4211949	8.59641689	9.0998933	9.0166815	9.4192907	8.8057534	8.7653492	8.8735243	9.2363731
06:15	8.95985792	9.3309729	9.85575301	9.279566	10.104889	9.345176	9.4874587	9.2547549	9.4946148	9.9848154
06:20	9.89000604	9.0255448	9.84413179	9.1261224	9.7815448	9.6225152	9.4700149	9.8596206	9.3283838	10.049282
06:25	9.84833332	8.3811643	8.59628195	9.2111328	9.2432871	8.9132438	8.9077773	9.3332711	8.6327449	8.9694776
06:30	10.1994666	9.40469	9.33290637	9.329851	9.1684081	9.2372212	9.2284952	9.0118983	9.5644223	9.5351283
06:35	9.39194646	8.8492123	9.1329706	9.0007269	8.6621113	8.9667837	9.1878411	8.9676542	8.7737337	8.6293851
06:40	9.00405271	9.1353157	8.96775749	9.214213	9.0849357	9.254756	8.9796045	9.5006997	9.232257	9.4847506
06:45	8.70426011	9.1787709	9.0526335	9.1454524	9.2001398	8.9002206	9.0785471	8.9699724	9.2296844	8.6236361
06:50	8.39083119	8.7180847	9.01011451	9.0642173	9.0314471	8.9160512	9.1664663	8.8676962	8.9224375	8.8988756
06:55	9.11071747	9.0778443	9.0973887	9.5655954	8.9362792	8.9932573	9.0824309	9.1274652	8.8778779	9.2382172
07:00	9.15542889	9.3178737	9.017456	9.3484465	9.2672875	9.3203664	9.3618507	8.9828771	9.1222761	9.3364023
07:05	9.56520072	9.1235063	8.98947614	9.3190013	8.9110416	8.8997338	9.248646	9.4075378	9.5017975	9.0733568
07:10	9.82950285	9.3587071	9.35263707	9.5790488	9.2341463	9.4561868	9.6199583	9.5676467	9.6372696	9.6044824
07:15	8.57047616	9.18389	9.36127892	9.9816064	9.2967176	9.4017296	9.7691452	9.6415729	9.6412266	9.3005002
07:20	9.82704456	9.6103267	9.51279309	10.094196	9.350508	9.4773494	9.9085234	10.252626	10.084995	9.259563
07:25	8.72330189	10.109665	9.34199421	9.8829716	9.4183778	9.6691403	9.7877834	10.041976	9.9815557	9.323795

Figure 0-8 Travel Time Matrix (TTM) file of loop data

7.4.2 User Defined Trips

Besides pre-defined trips, the DRIVE Net system also provide a function that allows users to select customized corridors to analyze travel time. Figure 7-9 shows the interface of the travel time analysis module using user-defined corridors. Users have two options to specify a corridor, shown in the upper red box in Figure 7-9. One is to choose the start point and end point from the map, which will generate a shortest path (route) between the two chose points. This method is implemented on the basis of the routing function of PostgreSQL. The other option is to input the start route and milepost and the end route and milepost, which will also generate a shortest path connecting the two input mileposts. Users can also combine these two functions to specify a corridor, such as choosing a start point on the map and inputting the end route and end milepost. To choose a point on the map, users click on the Show/Reset Loop Network button first and also check the checkbox for choosing a start/end point from the map. A customized date range, time

range and weekdays can also be specified in this function, shown in Figure 7-9. After choosing a corridor, users can preview the customized corridor by clicking on the Preview Customized Corridors button, shown in Figure 7-9. When the Travel Time Statistics button is clicked, the travel time analysis results appear in a table in a sub-window, which also contains an export button to export (download) the results into an Excel file, shown in Figure 7-10. The contents of the travel time analysis table are same as those of the pre-defined travel time analysis function, except that it can calculate the commute congestion cost (see Figure 7-11).

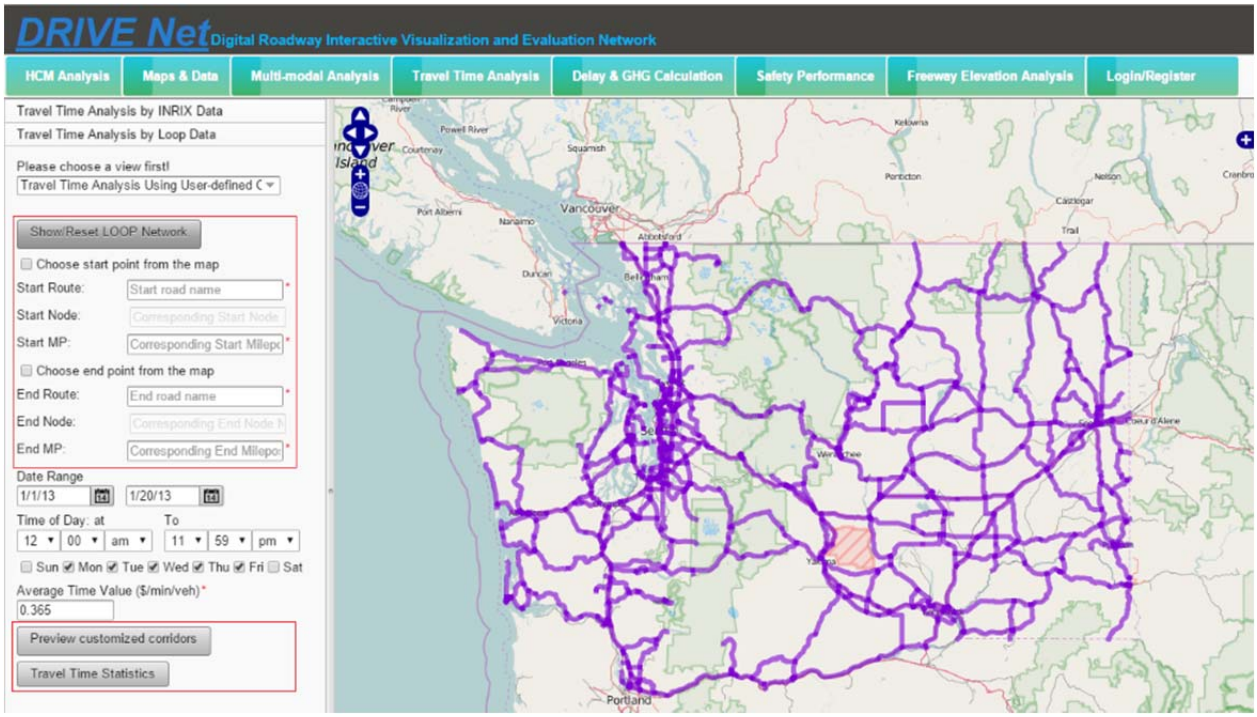


Figure 0-9 Travel time analysis using user-defined corridors

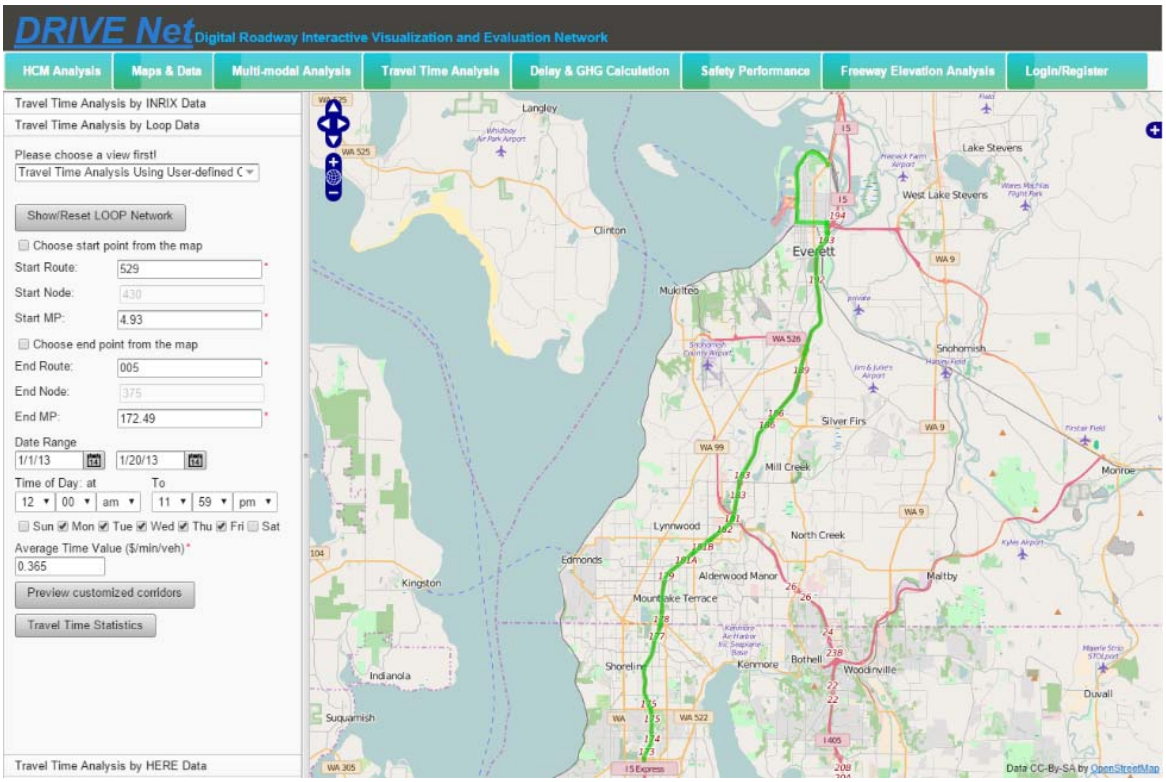


Figure 0-10 Preview customized corridors

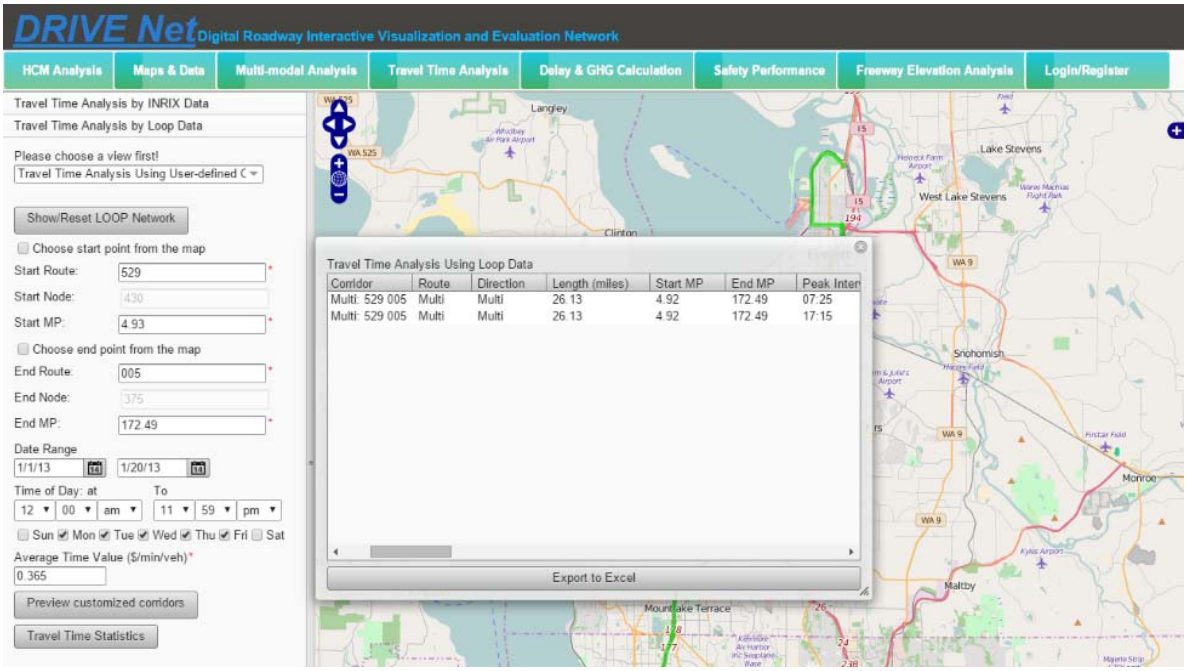


Figure 0-11 Travel time analysis of a user defined corridor

7.5 Travel Time Analysis with INRIX and HERE Data

The travel time analysis sections that use INRIX data and HERE data are very similar to the travel time analysis sections that use loop data. The differences are as follows:

- The pre-defined corridors for INRIX data, for which 12 corridors are defined, are different from those for loop data, as shown in the red box in Figure 7-12.
- Travel time analysis with HERE data does not have pre-defined corridor, as shown in Figure 7-13.

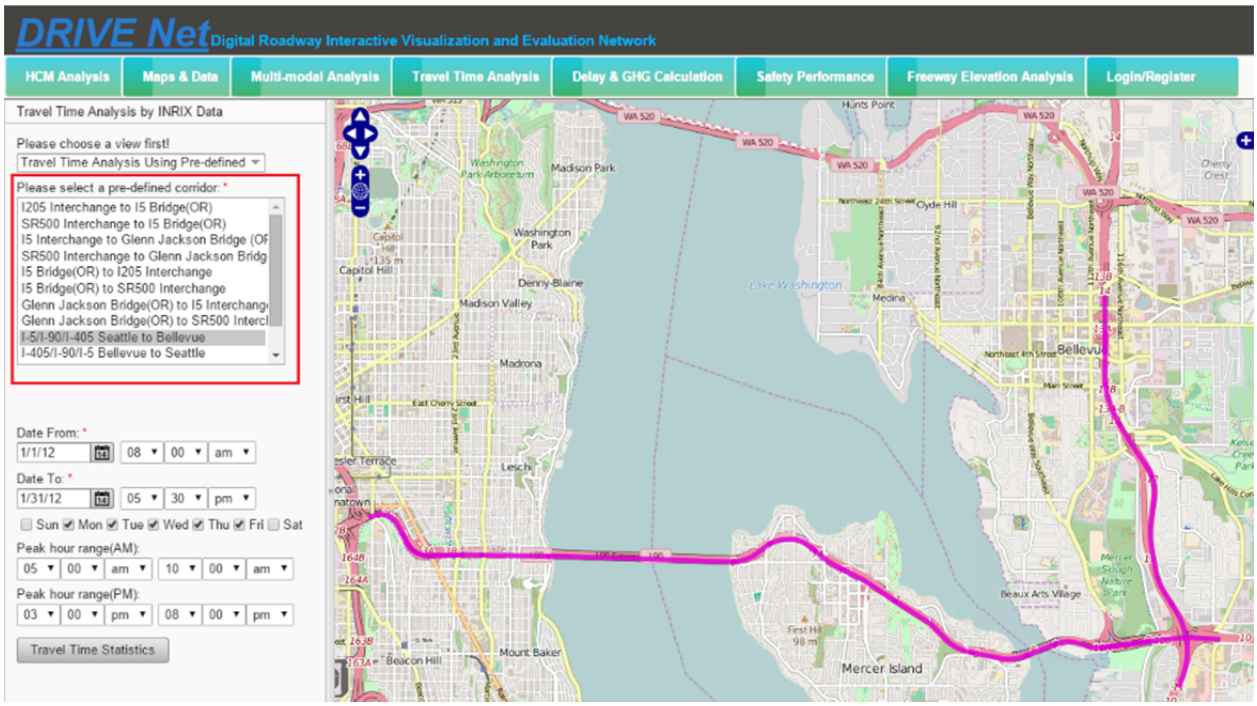


Figure 0-12 Travel time analysis with INRIX data

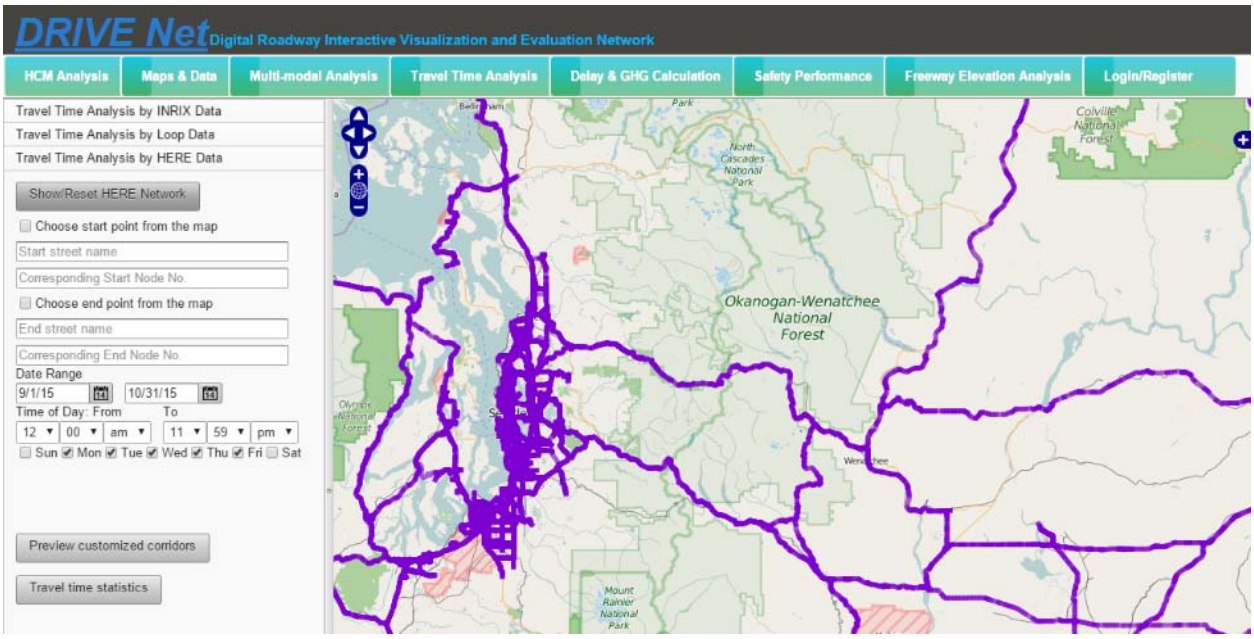


Figure 0-13 Travel time analysis with HERE data

7.6 Throughput Productivity

Throughput productivity can be computed for each travel direction at eight locations in the Puget Sound region, for a total of 16 locations. WSDOT measures throughput productivity by using the difference between the highest observed flow rate for that road section and the flow rate when the traffic speed falls below the maximum throughput speed (i.e., under-congested conditions). DRIVE Net computes this quantity as the throughput ratio, or the ratio of the current throughput performance to the maximum throughput. The maximum throughput speed is a user input field, but in general it should be set to the speed at which the highest 5-minute volume for the year was observed. When the traffic speed is above the maximum throughput speed, it is assumed that there is no loss in performance, and the throughput ratio is equal to 1.

To estimate throughput productivity in DRIVE Net, users start by selecting a location, shown in Figure 7-14. This highlights the selected location on the map. Next, users select an analysis year and maximum throughput speed, also shown in Figure 7-14. Finally, users click the Graph and Statistics button to display a throughput productivity summary for the selected location and analysis year. Figure 7-15 shows the summary statistics and throughput productivity plot for a location on NB I-5 at S 188th Street (MP 153.0). The summary graph shows throughput

performance for the highest observed 5-minute traffic volume for the selected year, as shown in Figure 7-16. Results can be exported as a Microsoft Excel® file, shown in Figure 7-17, by pressing the Export to Excel button on the screen..

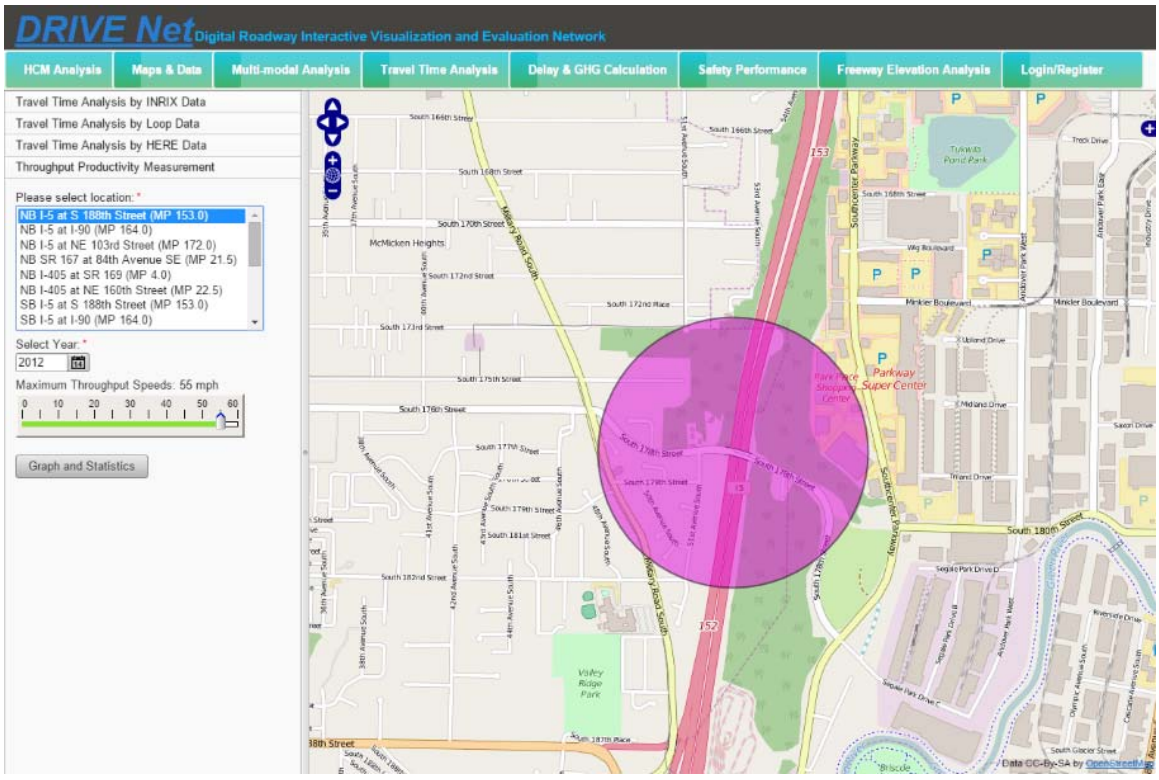


Figure 0-14 Throughput productivity measurement

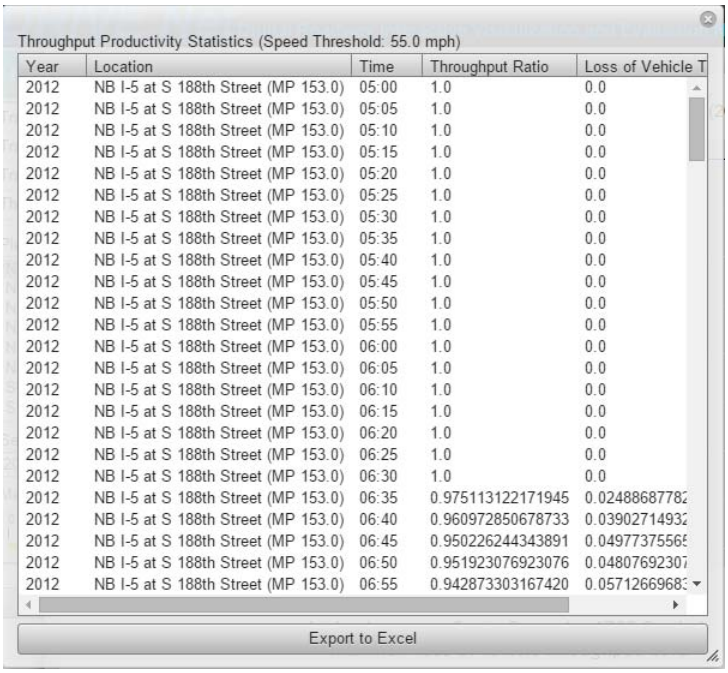


Figure 0-15 Throughput productivity statistics

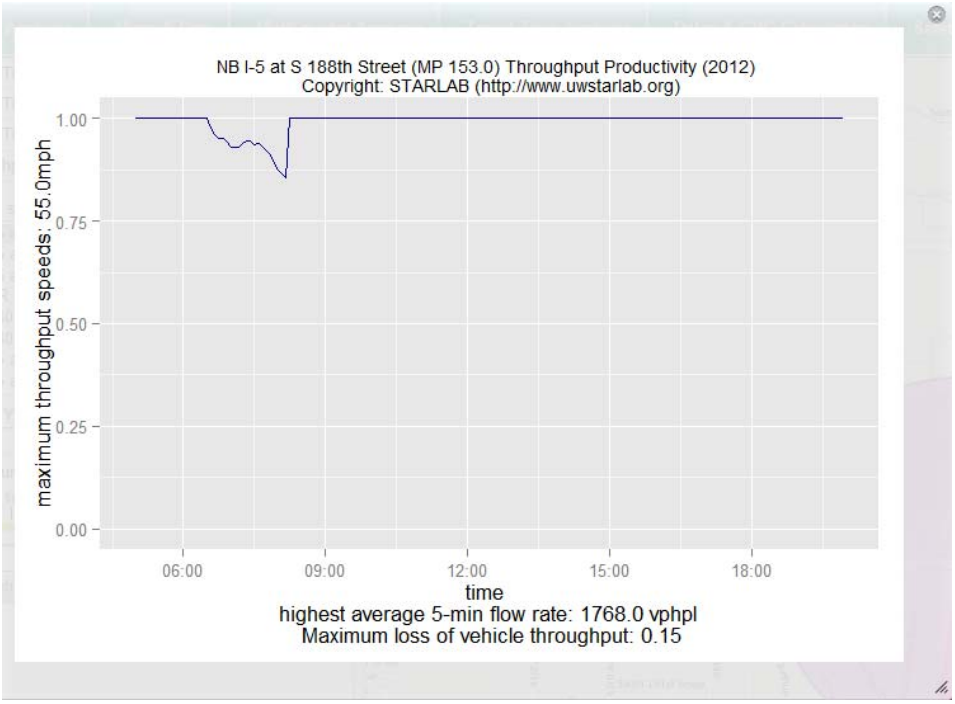


Figure 0-16 Throughput productivity plot

Throughput Productivity Statistics (Speed Threshold: 55.0 mph)

Year	Location	Time	Throughput Ratio	Loss of Vehicle Throughtput
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:00	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:05	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:10	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:15	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:20	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:25	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:30	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:35	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:40	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:45	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:50	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	05:55	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:00	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:05	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:10	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:15	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:20	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:25	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:30	1.00	0.00
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:35	0.98	0.02
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:40	0.96	0.04
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:45	0.95	0.05
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:50	0.95	0.05
2012.00	NB I-5 at S 188th Street (MP 153.0)	06:55	0.94	0.06
2012.00	NB I-5 at S 188th Street (MP 153.0)	07:00	0.93	0.07

Figure 0-17 Throughput productivity output to an Excel file

Chapter 8 HOV Analysis

8.1 Introduction

High occupancy vehicle (HOV) lanes are designed for faster traffic speeds, higher commuting efficiency, and better environmental -impacts than general purpose (GP) lanes (Washington State Department of Transportation, 2014). Because one prerequisite for higher traffic efficiency on HOV lanes is continuous delay on GP lanes, it is questionable whether constructing HOV lanes can marginally decrease delay in comparison with creating GP lanes (Dahlgren, 2001). Therefore, being able to measure and compare performance between GP and HOV lanes is essential for evaluating the actual effects of HOV lanes. In addition, freeway travel time reliability is critical for transportation system users because it is a key performance measure of a transportation system. So evaluating freeway travel time reliability on HOV lanes when incidents occur on GP lanes is also important.

8.2 Methodology for Analyzing the Travel Time Reliability of GP and HOV lanes—Percentile-Based Measures

To analyze travel time reliability on GP and HOV lanes, the percentile-based methodologies for measurement proposed by Tu et al. (2007) are employed. These percentile-based measures are widely used by researchers to evaluate travel time reliability and usually appear as "variability" measures. With these percentile-based measures, travel time reliability for each of four traffic conditions on HOV and GP lanes were obtained in this study. The four traffic conditions considered were normal condition with no incidents, incidents on the shoulder, incidents in a single lane, and incidents in multiple lanes. Travel time variability is reflected by the difference between the 90th (TT90) and 10th (TT10) percentile travel time in traffic condition k through Equation 8-1. The bigger that travel time variability is, the smaller that travel time reliability is, which means there is more travel time uncertainty.

$$\mathrm{TTV}_k = TT90_k - TT10_k \tag{8-1}$$

where,

TTV_k is travel time reliability (seconds/mile) in traffic condition k;

$TT90_k$ is 90th percentile travel time per unit road distance (seconds/mile) in traffic condition k;

$TT10_k$ is 10th percentile travel time per unit road distance (seconds/mile) in traffic condition k;

k is the normal condition with no incidents, incidents on the shoulder, incidents in a single lane, and incidents in multiple lanes.

8.3 Results and Discussion

Travel time reliability on Interstate 405 North, Interstate 405 South, and Interstate 5 South was examined. . Interstate 5 North was excluded from the analysis because of incomplete loop detector data from the HOV lane. Travel time data for the three routes were collected under four conditions. The four conditions were normal conditions with no incidents, incidents on the shoulder, incidents in a single lane, and incidents in multiple lanes. Figures 9, 10, 11, and 12 show comparisons of percentile travel time and travel time variability between HOV and GP lanes on I-5 South, I-405 North, and I-405 South.

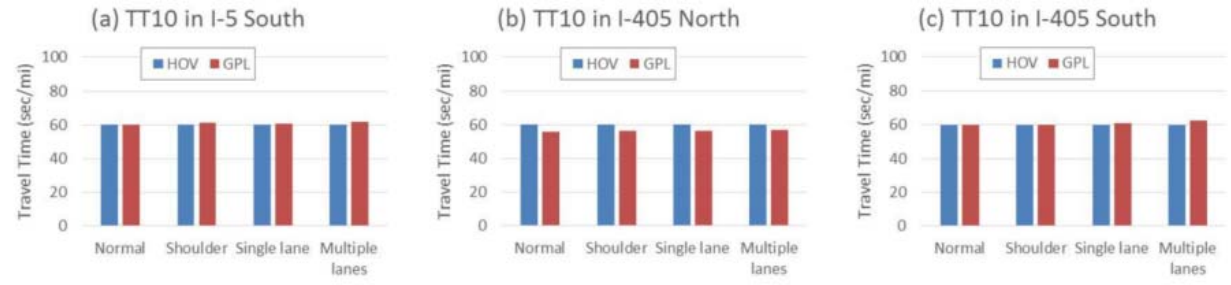


Figure 0-1 Comparison of 10th percentile travel time between HOV and GP lanes in I-5 South, I-405 North, and I-405 South

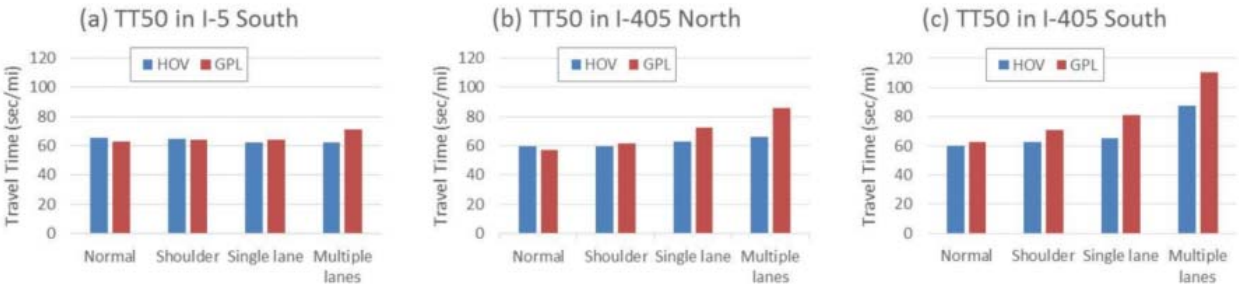


Figure 0-2 Comparison of 50th percentile travel time between HOV lanes and GP lanes on I-5 South, I-405 North, and I-405 South

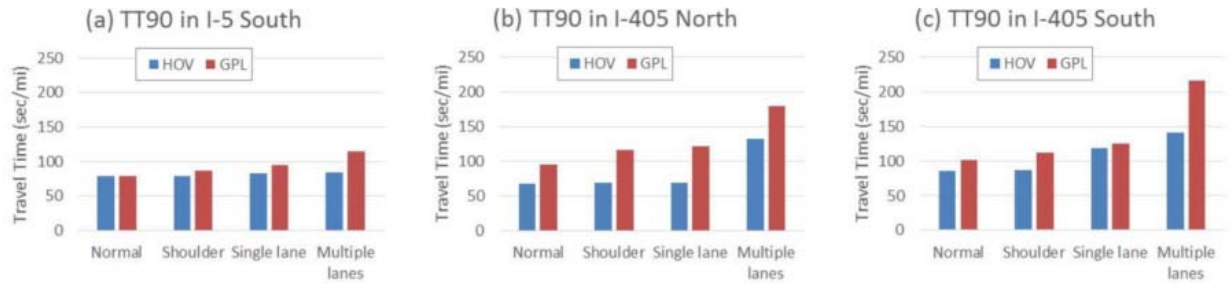


Figure 0-3 Comparison of 90th percentile travel time between HOV lanes and GP lanes on I-5 South, I-405 North, and I-405 South

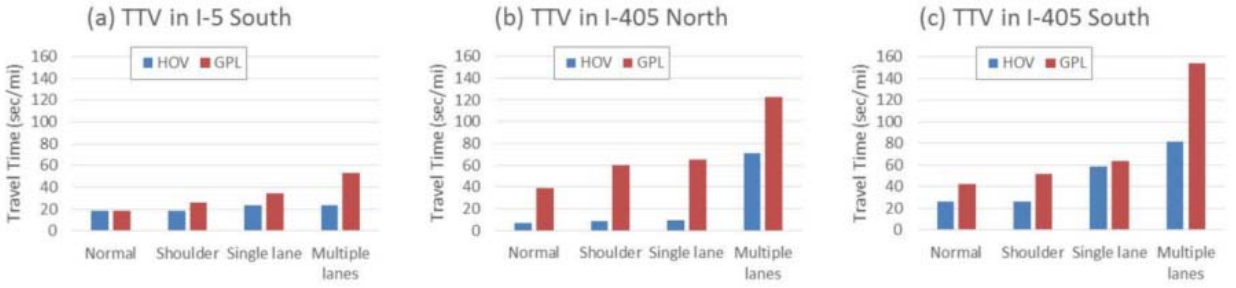


Figure 0-4 Comparison of travel time variability between HOV lanes and GP lanes on I-5 South, I-405 North, and I-405 South.

8.3 Impacts of Shoulder Incidents

The TT10 and TT50 values for HOV and GP lanes on each route were almost the same under the shoulder incident condition, whereas the TT90 values for HOV and GP lanes were very different, with results of 9.98 percent on I-5 South, 40.81 percent on I-405 North, and 22.74 percent on I-405 South. TT10 and TT50 values represent low traffic volume conditions. TT90 represents the travel time values observed in the peak hours, and freeway bottlenecks can be easily created by

shoulder incidents. Therefore, the differences between the TT90 values for HOV and GP lanes are remarkable, although shoulder incidents are minor incidents. The reason maybe that when GP lanes become congested as the result of an incident, the single-occupant vehicles cannot move to the adjoining HOV lane(s). Therefore, the traffic incident has only a slight impact on the HOV lanes. On the other hand, if an incident happens in the HOV lane, those drivers who were in the HOV lane can shift to the GP lanes to avoid congestion. The differences between travel time reliability (TTV) for HOV and GP lanes under shoulder incident conditions were 29.37 percent on I-5 South, 85.36 percent on I-405 North, and 48.54 percent on I-405 South. Thus HOV lanes had better travel time reliability than GP lanes.

8.4 Impacts of Single Lane Incidents

The TT10 values in HOV lanes and GP lanes were almost the same on all the three routes under the single lane incident condition. However, travel time differences between HOV and GP lanes were observed in both TT50 and TT90 values, which differed from the shoulder incident condition. The differences in TT50 values were 2.68 percent on I-5 South, 13.47 percent on I-405 North, and 20.02 percent on I-405 South, while the difference in TT90 values were 12.72 percent on I-5 South, 42.96 percent on I-405 North, and 4.75 percent on I-405 South. The reason may be that because single lane incidents can block one of the travel lanes on a freeway road segment and can thus induce a bottleneck at a relatively low traffic flow rate, this incident type can have a more obvious impact on travel time than shoulder incidents. The TTV differences between HOV and GP lanes for single lane incidents were 32.74 percent on I-5 South, 85.91 percent on I-405 North, and 7.66 percent on I-405 South. Therefore, under the impacts of single lane incidents, HOV lanes may have a better travel time reliability than GP lanes.

8.5 Impacts of Multiple Lanes Incidents

The differences between TT10 values for HOV and GP lanes was slight under the impacts of multiple lanes incidents, whereas the differences were noticeable for TT50 and TT90 values. The differences for TT50 values were 11.74 percent on I-5 South, 23.24 percent ion I-405 North, and 20.61 percent on I-405 South, while the differences for TT90 were 26.88 percent on I-5 South, 26.86 percent on I-405 North, and 34.48 percent on I-405 South. These differences were larger

than those for single lane incidents. This may be because multiple lane incidents are the most serious incident type and can block multiple lanes. Therefore, the threshold traffic flow value for traffic breakdown is smaller for multiple lane incidents than for single lane and shoulder incidents. HOV and GP lanes are both influenced heavily by multiple lane incidents, and the situation becomes worse with an increase in traffic volumes. However, even for multiple lane incidents, travel times in HOV lanes are shorter than those in GP lanes, since the single occupant vehicles are not allowed to use the HOV lanes.

Under multiple lane incident circumstances, the TTV differences between HOV and GP lanes were 55.36 percent on I-5 South, 41.8percent on I-405 North, and 46.9percent on I-405 South, meaning that HOV lanes had better travel time reliability.

8.6 Comparison of the Impacts of Different Incidents

All the incident types had a negative influence on travel time reliability. Nevertheless, the levels of impact were different. Among all the incident types, shoulder incident conditions resulted in the least affected travel time reliability, single lane incident conditions ranked second, and multiple lane incident conditions caused the worst travel time reliability. As expected, the more serious the incidents were, the less reliable travel time was. The difference between TTV values for HOV and GP lanes on each route also varied under different incident types. For example, on I-5 South, it was -0.87 percent under normal conditions, 29.37 percent with shoulder incidents, 32.74 percent with single lane incidents, and 55.36 percent with multiple lane incidents. This shows that when incidents occur on GP lanes, the traffic in HOV lanes will not be significantly affected and that HOV lanes generally provide much better travel time reliability.

8.7 Comparison of I-5 and I-405

The data for I-405 South and I-405 North were averaged to compare with those for I-5 South. This showed that I-5 GP lanes had better travel time reliability than I-405 GP lanes, since the travel time variability indexes for I-405 were larger than those for I-5 overall. However, in the HOV lanes, I-5 showed higher reliability with single lane incidents and multiple lanes incidents, while the travel time reliabilities of I-5 and I-405, on average, were similar under normal and shoulder incident conditions. For GP lanes, the possible explanation is that I-5 has three general

purpose lanes while I-405 has only two. Therefore, I-5 is more reliable than I-405 under normal conditions because vehicles in I-5 have more lane options to shorten their travel time. In addition, once incidents happen, vehicles on I-5 have more flexibility to choose lanes to avoid the queue. In terms of HOV lanes, the travel time reliabilities under normal and shoulder incident conditions, the least impactful incident type, were similar, since both I-5 and I-405 have only one HOV lane. Nevertheless, when more serious incident types occurred, the advantage of more GP lanes on I-5 resulted in better travel time reliability. To sum up, roadway geometry does influence travel time reliability. More general purpose lanes can result in higher travel time reliability on GP lanes when traffic incidents occur.

8.8 Conclusions

Three incident types (i.e., shoulder incident, single lane incident, and multiple lane incident) were considered for analysis. Travel time reliability on GP and HOV lanes were analyzed by using a percentile-based measure. The preliminary results suggested that incidents result in lower values of travel time reliability. The results also showed that, of the three incident types, multiple lane incidents have the most significant impact on freeway route travel time variability, while shoulder incidents have the least impact. Generally, HOV lanes have better travel time reliability than GP lanes under the same incident conditions.

Chapter 9 Travel Delay Calculation

9.1 Introduction

Travel delay is defined as the amount of extra time travelers spend in traffic as a result of increased traffic volumes. Travel delay can be calculated for the number of vehicles or people on the road and is measured relative to a speed threshold such as maximum throughput speed or posted speed. Traffic delays may also be caused by a reduction in the number of lanes or a lower speed limit. Generally, traffic delays may be calculated for both delays under uncongested conditions and congested conditions. It is important to accurately estimate travel delay not only to help travelers re-plan their travel routes in response to real-time delay calculation results, but also to help transportation agencies design effective and efficient traffic management and control strategies.

Congestion has been a major traffic operational problem in the Puget Sound region for the past decade. It is mainly a consequence of significant increases in traffic demand in the metropolitan areas over the past 25 years. Total vehicle miles traveled (VMT) increased by 78.5 percent from 1980 to 1992 and 23.3 percent from 1992 to 2004 (PSRC 2005). With this fast growth of VMT and limited resources for improving freeway infrastructure, traffic congestion continues to deteriorate. In 2003, the congested period on highways reached 7.6 hours per day (Dutzik and Pregulman 2003). This deterioration increases the urgency of mitigating traffic congestion in Washington State. In the past decade, WSDOT has invested in multiple traffic congestion mitigation projects (Hallenbeck et al. 2003). The main objectives of these projects have been to improve WSDOT's understanding of traffic congestion causes and impacts, and to identify the most effective countermeasures against such traffic congestion.

Recent studies have found that more than 50 percent of freeway congestion is the result of traffic incidents (Transportation Research Board 2000). Special attention should be paid to travel delays caused by incident-related congestion because of the fact that incident-induced congestion may be cost-effectively alleviated through traffic management, control, and incident response. To mitigate incident-induced delay (IID), a better understanding of incident impacts on traffic and traffic evolution during an incident is indispensable. Unfortunately, little work has

been completed to evaluate the impacts and causes of incidents in Washington's metropolitan areas. The Washington State Transportation Center (TRAC) developed an algorithm to identify and estimate incident-related congestion on the basis of loop occupancy data from existing loop detection systems (Hallenbeck et al. 2003). The loop occupancy profile extracted directly from loop detector data are compared with a background occupancy profile to identify the occurrence of an individual incident and to estimate its influence. The background occupancy profile is created from the medians of loop occupancy values collected on weekdays without incidents. This algorithm is straightforward to apply. However, test results have shown that this algorithm may not be sensitive enough to capture all incidents. A preliminary analysis found that about 50 percent of incidents were not detected by this loop occupancy-based algorithm. This may be due to the fact that its fixed background occupancy profile is unsuitable for traffic conditions that are significantly different from the ordinary scenarios represented by the median occupancy values. Furthermore, using only loop occupancy data cannot accurately represent true traffic conditions on freeways. For instance, high loop occupancies may be a result of either a few slow moving vehicles passing over a loop or many high speed vehicles flowing over a loop. Consequently, a better algorithm for estimating incident-induced delay needs to be developed.

9.2 Background

Traffic incidents result in remarkable travel delays on freeways. To minimize the impacts of traffic incidents, researchers have spent enormous effort developing procedures to detect the occurrence of incidents. Methods used for incident detection include artificial neural networks (Ritchie and Cheu 1993), a loop occupancy-based approach (Lin and Daganzo 1997), and wavelet technique (Teng and Qi 2003). The application of these methods helps shorten the time needed for incident detection and hence reduces incident response time, which consequently lowers incident impacts on traffic movements.

However, those incident detection procedures do not provide information about the impacts of the incidents on traffic congestion. This has led to an interest in developing procedures for estimating IID. Existing procedures for calculating IID have been based on either the deterministic queuing theory or shock wave analysis. The queuing theory-based procedures calculate IID by using a queuing diagram formed by cumulative vehicle arrival and departure

curves. The area between these curves represents the total delay in units of vehicle-hours. The queuing diagram approach was also employed by Cohen and Southworth (1999). They proposed a simple model for estimating the mean and variance of time lost as a result of incidents on freeways. Olmstead (1999) showed that the queuing model may underestimate the total delay if the model assumes that the delay due to an average incident is the same as the average delay due to incidents. Queuing theory has also been applied to estimate delays at work zones on freeways (Chien and Chodhury 2000). Li et al. (2006) recently introduced an incident duration model and a reduced capacity model for use with the queuing theory to estimate IID on freeways. Their delay estimation model provides reasonable estimates of the mean as well as the variance of IID. Traffic flow has some characteristics similar to those of fluid flow. Therefore, several researchers have attempted to use kinematic wave theory to explain the behaviors of traffic flow. These attempts have led to the development and application of shock wave analysis for estimating IID.

Hallenbeck et al. (2003) studied the nature and cause of traffic congestion on freeways in Seattle's metropolitan area. The occurrence and duration of traffic congestion caused by incidents were identified by comparing the traffic profile of lane occupancy on a day with incidents with a background occupancy profile that represented the typical traffic condition for incident-free days. The difference between the two profiles was used to calculate the delay caused by incidents. However, the test results from the process included traffic congestion that sometimes moved from upstream to downstream locations, which could be questionable. Nonetheless, this study built a solid foundation for further studies on incident detection and delay estimation in Washington state.

9.3 Methodology

In DRIVE Net, two quantified indexes are provided to evaluate travel delay and delay cost: total vehicle hours delay and total cost. First, the hours of travel delay can be defined as:

$$H = (M/S) - (M/S_T) \tag{9-1}$$

where H means hours of travel delay, M represents vehicle miles traveled, S is the travel speed, and S_T indicates travel threshold speed. According to WSDOT delay calculation methodology,

maximum throughput speed (85% of posted speed limit) is used as the threshold in order to measure delay relative to a highway's most efficient operating condition. Travel speed is averaged hourly for each weekday by highway segment. Any of the 120 speed data points (5 days × 24 hours) that show speeds slower than the threshold speed are identified as "experiencing delay." Accordingly, one can further calculate annual hours of vehicle delay (AHD), which is all travel delay, reported in vehicle hours, experienced for the year. This delay is also summarized by urban area and for selected major commute corridors, and it is defined as:

$$\mathrm{AHD} = \sum_{i=1}^{n} (M_i / S_i) - (M_i / S_T) \quad (9\text{-}2)$$

where i means the ith weekday in a year, n is the total number of weekdays in a year, and correspondingly, M_i and S_i are the vehicle miles traveled and travel speed in the ith weekday of the year.

For the cost caused by delay, annual cost of vehicle delay is used to estimate economic impact. Its calculation is based on lost productive time, wasted fuel, and additional vehicle maintenance costs due to extra time spent in traffic. The cost of vehicle delay is calculated by applying monetary values to the estimated hours of delay incurred by passenger and truck travel, plus additional vehicle operating costs:

$$C = (C_H \times \mathrm{CPI}_{change}) \times H \quad (9\text{-}3)$$

where C is the annual cost of vehicle delay, its unit is dollars, CPI means Consumer Price Index and CPI_{change} represents the difference of CPI between two years, and H is the hours of travel delay. On the basis of WSDOT research, the value of time for passenger trips is assumed to be half of the average wage rate, while it is assumed to be 100 percent of the wage rate plus fringe benefits for truck drivers. The Consumer Price Index (CPI) from the Bureau of Labor Statistics is applied to these values to reflect the influence of inflation. Furthermore, WSDOT's Urban Planning Office recommends using $26.52/hour of delay (in 2013 dollars) as parameter when calculating the cost of delay for traffic that is a mixture of vehicle types.

9.4 Results

DRIVE Net provides the results of delay calculation for specific routes, which can be based on pre-defined corridors or selected by the user. Figure 9-1 shows the calculation of delay and its cost for the corridor I-405 Bellevue to Lynnwood. The data used for delay calculation were collected in 2012. Users can also regulate and set the values for the posted speed limit, travel costs per hour, and CPI. For this corridor, the total delay in 2012 was about 1,710,669 vehicle hours and a total cost of about $52,353,463.

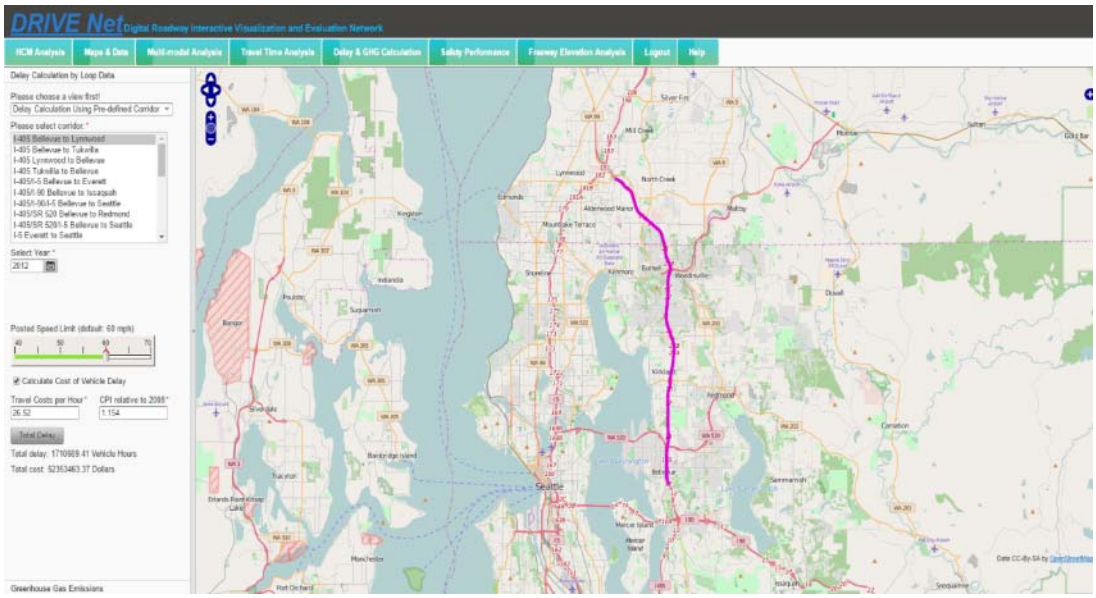


Figure 0-1 Delay calculation in a pre-defined corridor

There are two ways for users to select corridors. First, users can load available loop detector networks and then select the start point and end point in the network (see Figure 9-2). The corridor selected by the user can be previewed as in Figure 9-3

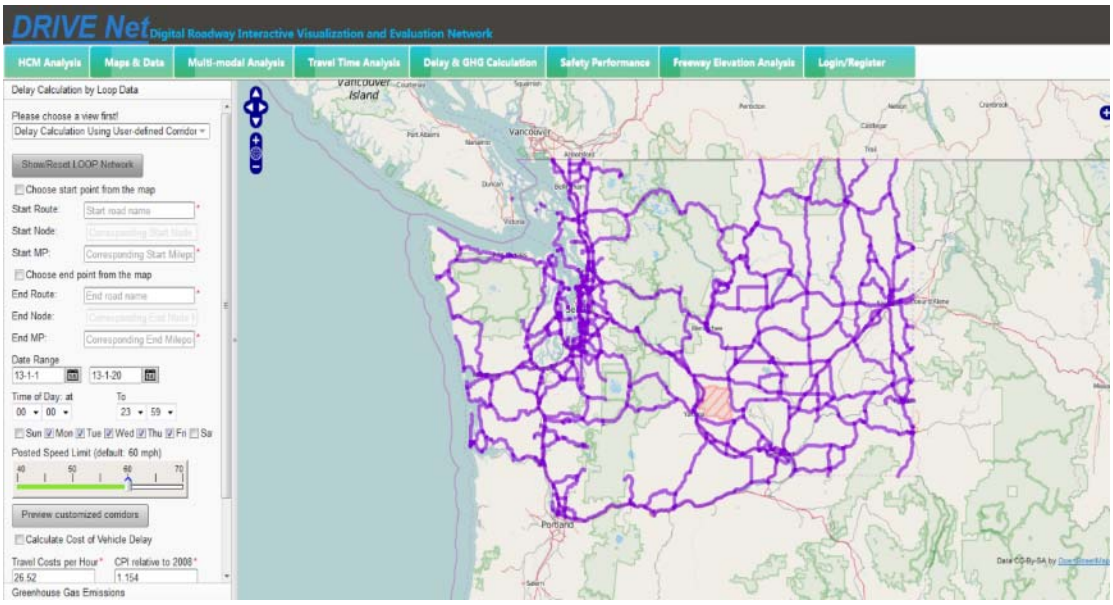


Figure 0-2 Loop detector networks

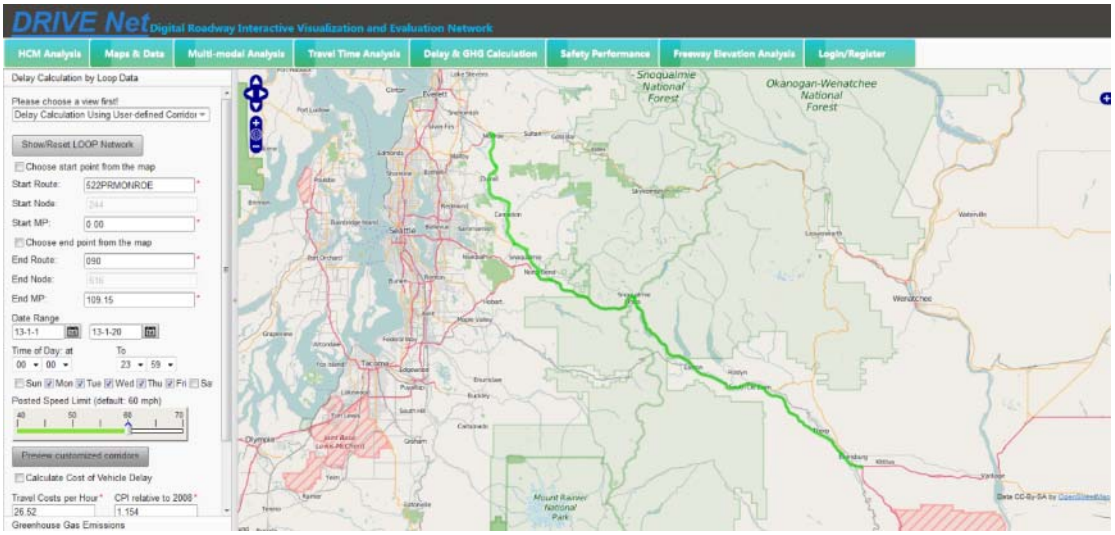


Figure 0-3 Preview of a selected corridor

The second way is to directly enter the start route, start milepost, end route, and end milepost in the module, and then the delay and cost can be calculated. Figure 9-4 shows the calculation process of delay and its cost for a corridor selected by the user. In this module, users first choose the start and end routes; for example, the route starts from 104 to 190. Then the user sets the start milepost and end milepost, 29.14 to start and 0.53 to end. Next, the user can choose the calculation time period, including dates, hours, and weekdays or weekends. The green line in

Figure 9-4 is a preview of the selected corridor. Similarly, users can also regulate and set the values for the posted speed limit, travel costs per hour, and CPI. The results for this corridor show that the total delay during the selected time period is about 597,542 vehicle hours, with a total cost of about $18,287,230.

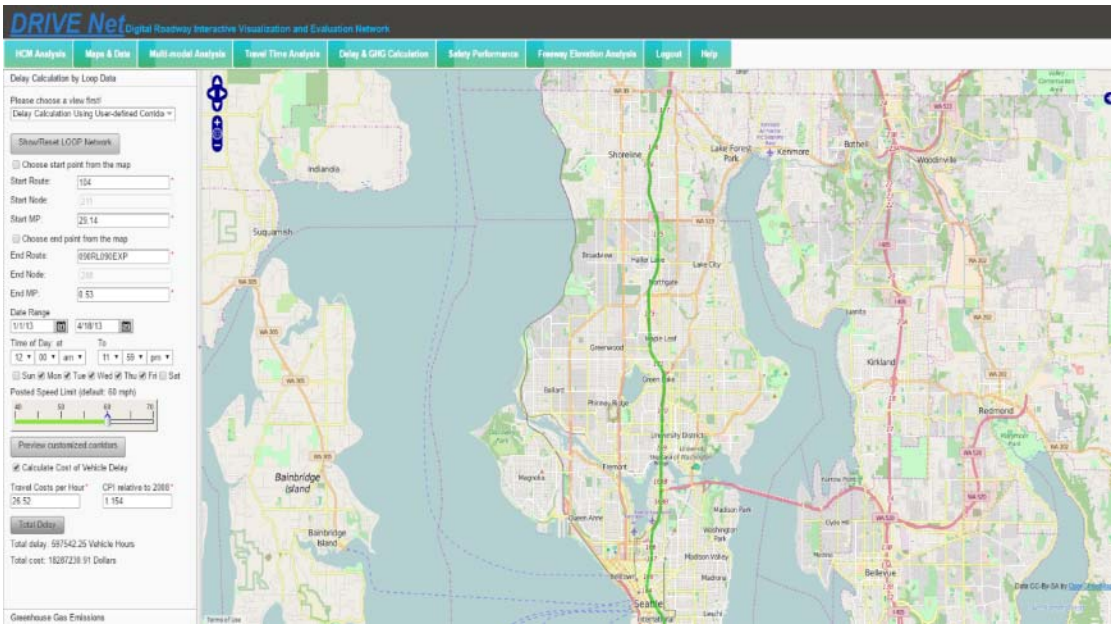


Figure 0-4 Delay calculation in a user-selected route

Chapter 10 Safety Analysis

10.1 Introduction

For the purposes of prioritizing safety improvements on the roadway network, identifying sites with a consistently elevated accident risk, often referred to as hotspots or black spots, is of critical importance. To address this need, a number of analytical methods for hotspot identification (HSID) have been developed over the last several decades, with the overarching objective of optimizing the allocation of limited funding. An inaccurate HSID method will result in inefficient allocation of safety treatment resources, with potentially serious costs in terms of overall safety performance of the network. The need for accurate methods to identify and prioritize accident prone locations is underscored by the U.S. 2012 Federal Moving Ahead for Progress in the 21st Century Act (MAP-21) (Federal Highway Administration U.S. Department of Transportation, 2014), which emphasizes data-driven crash risk analysis and safety treatment prioritization. Furthermore, the performance reporting requirements outlined in MAP-21 provide an additional layer of incentive for public agencies to maximize the impacts of safety spending by selecting and treating sites with high improvement potential.

A number of studies in past years have focused on the accident frequency- (AF) or accident rate- (AR) based HSID methods, which rely on observed accident counts as the primary measure of accident risk. Because sites are ranked and identified on the basis of observed accident data only, there is no mechanism for identifying sites with elevated risk (due to some combination of geometric and traffic characteristics) but few accidents. Furthermore, these methods cannot distinguish between actual high risk locations and those with a higher occurrence of accidents due to random fluctuations. The Empirical Bayes (EB) HSID method addresses these issues by combining two clues, the historical crash record of the entity and the expected number of crashes obtained from a safety performance function (SPF) for similar entities. This approach is less sensitive to random fluctuations in accident frequency, and in theory it can identify truly high risk locations with greater accuracy. Building on the EB approach, additional methods have been developed that are based on estimated accident reduction potential (ARP). Such methods attempt to quantify the difference between the actual accident count at the location of interest (as estimated with the EB method) and the expected

accident count for similar locations, under the supposition that this difference represents the potential for improvement.

10.2 Background

10.2.1 Hotspot Identification Methods

AF-based HSID methods have been used for many years. Such approaches typically rank locations or segments along a highway by observed accident count over a specified time interval and define hotspots as those whose counts exceed some critical value (Deacon et al., 1975). Road segments or intersections are ranked by accident count among similar locations (such as along a relatively homogeneous section of highway) to ensure that the identified hotspots represent specific opportunities for remediation instead of some inherent characteristic of a particular roadway class or driver population. One criticism often raised with regard to the AF method is that this approach lacks the ability to differentiate between actual hotspots and locations with increased accident frequency attributable to the randomness of traffic accidents (Deacon et al., 1975, Cheng and Washington, 2008).

It is readily apparent that, all else being equal, a segment with higher traffic volume can be expected to have a higher accident count, and so hotspots identified with AF methods tend to over-represent high volume locations that may or may not be amenable to remediation efforts (Hauer, 1996). In response, AR methods have been developed that rely on accident count per unit of traffic volume for HSID, typically in units of accidents per million vehicle miles traveled. As with the AF methodology, sites are ranked by accident rate, and those with rates that exceed a critical value are identified as hotspots. Implicit in this approach is the assumption that accident count and exposure are linearly related, which is often not the case. In addition, by normalizing accident count by entering traffic volume, locations with very low traffic volume are sometimes over represented (Hauer, 1996, Persaud et al., 1999).

The EB method for traffic accident HSID was introduced by Abbess et al. (1981) to address issues with existing methodologies, most notably regression-to-the-mean (RTM) bias and low precision due to limited accident history. It has since been refined and widely used in a range of safety performance modeling applications (Cao et al., 2012; Mountain et al., 1996; Zou

et al., 2015). In the EB crash modeling procedure, the expected number of crashes at a location is estimated by combining two pieces of information: (1) the accident count at the location of interest and (2) the expected accident count at locations determined to be similar on the basis of traffic and roadway characteristics (Hauer et al., 2002). The method assumes that the actual accident count for the location of interest is available, and the expected accident count for similar locations is generally estimated from the safety performance function (SPF). The SPF that describes accident counts as a function of traffic volume, lane width, etc. is typically fitted by using the negative binomial (NB) regression model. The observed accident count for a given roadway segment is combined with the expected value estimate as shown in Eq. (10-1).

$$\hat{\hat{\mu}}_i = w_i \hat{\mu}_i + (1 - w_i) y_i \tag{10-1}$$

where,

$\hat{\hat{\mu}}_i$ = the EB estimate of the expected number of crashes per year for site i;

$\hat{\mu}_i$ = the estimated number of crashes per year by the SPF for given site i (estimated using a NB model);

$w_i = \dfrac{1}{1 + \alpha \hat{\mu}_i}$ is the weight factor estimated as a function of $\hat{\mu}_i$ and dispersion parameter α; and

y_i = the observed number of crashes per year at site i.

Another measure often used in HSID is accident reduction potential (ARP). Research originally suggested that ARP be estimated as the difference between the observed accident count at the site of interest and the expected count as estimated from a set of reference sites. More recently, others have proposed that the observed accident count at the site of interest be replaced by the EB estimated accident count. This approach can account for random fluctuations in accident frequency and so give a better estimate of the true safety of the location of interest. Using the EB estimated accident count, the ARP is calculated as shown in Eq. (10-2).

$$ARP_i = w_i \hat{\mu}_i + (1 - w_i) y_i - \hat{\mu}_i \quad (10\text{-}2)$$

where,

ARP_i = ARP for site i.

Persaud et al. (1999) suggested that a better estimate of the true ARP can be derived by using a full predictor set in the EB-estimated accident count and a subset of available regressors (i.e., those not describing a correctable, site-specific geometric feature) in the expected accident count model. This way, the estimated ARP is a measure of the difference between the EB-estimated "true" safety and the expected safety of what could be considered a base scenario.

10.2.2 Negative Binomial Model

For highway safety, the dispersion parameter of NB models refines the estimates of the predicted mean when the EB method is used. So far, transportation safety analysts use the NB distribution model most frequently to generate SPFs. The NB model has the following structure: the number of crashes y_i during some time period is assumed to be Poisson distributed, which is defined by:

$$p(y_i \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y_i!} \quad (10\text{-}3)$$

where,

λ = mean response of the observation.

The NB distribution can be viewed as a mixture of Poisson distributions where the Poisson rate is gamma distributed. The probability density function (PDF) of the NB is defined as follows:

$$f(y_i \mid \mu, \alpha) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i + 1)} (\frac{\alpha\mu}{1 + \alpha\mu})^{y_i} (\frac{1}{1 + \alpha\mu})^{1/\alpha} \quad (10\text{-}4)$$

where,

$\mu =$ mean response of the observation; and,

$\alpha =$ dispersion parameter.

In comparison to the Poisson distribution, the NB distribution can allow for over-dispersion.

10.3 Data Description

Data used for developing safety performance functions (SPFs) were obtained from the Highway Safety Information System (HSIS). The data consisted of roadway data and crash records. The roadway data contained segment length, AADT, number of lanes, lane width, left and right shoulder width, median width, speed limit, and weather conditions when the accident occurred. For each of the six years from 2006 to 2011, crash data were extracted for road segments. This project analyzed interstate freeways in urban areas because these segments had similar characteristics. Each segment had been examined, and there were no significant differences in these six years. The statistics of the road segments are shown in Table 10-1.

Table 0-1 Summary statistics of characteristics for individual urban interstate freeway segments in Washington state for years 2006 – 2011

	Year 2011			
Explanatory variables	Min.	Max.	Mean	S.D.
AADT	9174	229101	107415	54321.69
Lane width	11	22	12.2	1.01
Number of lanes	1	9	5.79	1.53
Left shoulder width	0	12	3.305	3.62
Right shoulder width	0	21	9.41	2.61
Median width	0	300	45.9	43.60
Design speed limit	50	70	62.18	4.19

Segment length	0.1	2.02	0.221	0.17
Total number of crashes	1	59	6.336	7.13
Number of crashes in normal weather conditions	0	44	4.247	5.05
Number of crashes in abnormal weather conditions*	0	33	2.089	2.71
	Year 2010			
Explanatory variables	Min.	Max.	Mean	S.D.
AADT	1492	237186	105706	55598.08
Lane width	11	26	12.29	1.36
Number of lanes	1	9	5.72	1.59
Left shoulder width	0	12	3.38	3.65
Right shoulder width	0	21	9.26	2.84
Median width	0	300	43.82	39.87
Design speed limit	50	70	61.7	4.08
Segment length	0.1	2.02	0.229	0.18
Total number of crashes	1	107	6.709	8.99
Number of crashes in normal weather conditions	0	76	4.291	6.24
Number of crashes in abnormal weather conditions	0	31	2.418	3.30
	Year 2009			
Explanatory variables	Min.	Max.	Mean	S.D.

AADT	2001	239371	105767	56381.9
Lane width	11	26	12.33	1.47
Number of lanes	1	9	5.713	1.61
Left shoulder width	0	12	3.515	3.67
Right shoulder width	0	21	9.23	2.85
Median width	0	300	46.13	42.71
Design speed limit	60	70	61.7	3.97
Segment length	0.1	2.02	0.2295	0.18
Total number of crashes	1	59	6.402	7.57
Number of crashes in normal weather conditions	0	44	4.557	5.66
Number of crashes in abnormal weather conditions	0	20	1.845	2.44
	Year 2008			
Explanatory variables	Min.	Max.	Mean	S.D.
AADT	1053	236206	108368	56144.44
Lane width	10	45	12.34	1.71
Number of lanes	1	9	5.704	1.59
Left shoulder width	0	12	3.407	3.64
Right shoulder width	0	21	9.279	2.78
Median width	0	300	45.27	43.19

Design speed limit	60	70	61.63	3.88
Segment length	0.1	2.02	0.2299	0.19
Total number of crashes	1	80	6.627	7.67
Number of crashes in normal weather conditions	0	59	4.242	5.41
Number of crashes in abnormal weather conditions	0	25	2.385	2.90
	Year 2007			
Explanatory variables	Min.	Max.	Mean	S.D.
AADT	1482	240266	111097	57537.4
Lane width	11	45	12.36	1.80
Number of lanes	1	9	5.674	1.59
Left shoulder width	0	12	3.496	3.64
Right shoulder width	0	21	9.272	2.82
Median width	0	300	46.28	43.26
Design speed limit	60	70	61.59	3.88
Segment length	0.1	2.02	0.2298	0.18
Total number of crashes	1	73	7.155	8.12
Number of crashes in normal weather conditions	0	53	4.806	5.92
Number of crashes in abnormal weather conditions	0	28	2.348	2.83
	Year 2006			

Explanatory variables	Min.	Max.	Mean	S.D.
AADT	982	282619	113233	56985.19
Lane width	11	26	12.28	1.24
Number of lanes	1	9	5.74	1.54
Left shoulder width	0	12	3.433	3.67
Right shoulder width	0	21	9.286	2.81
Median width	0	300	45.95	43.19
Design speed limit	60	70	61.54	3.84
Segment length	0.1	2.02	0.229	0.18
Total number of crashes	1	90	7.791	8.99
Number of crashes in normal weather conditions	0	69	5.238	6.30
Number of crashes in abnormal weather conditions	0	27	2.553	3.27

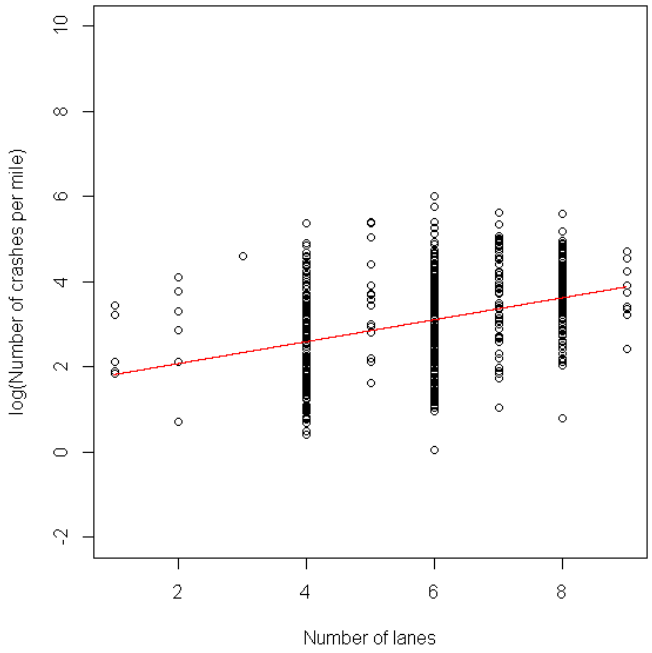
* Abnormal weather condition means rain or snow

10.4 Results

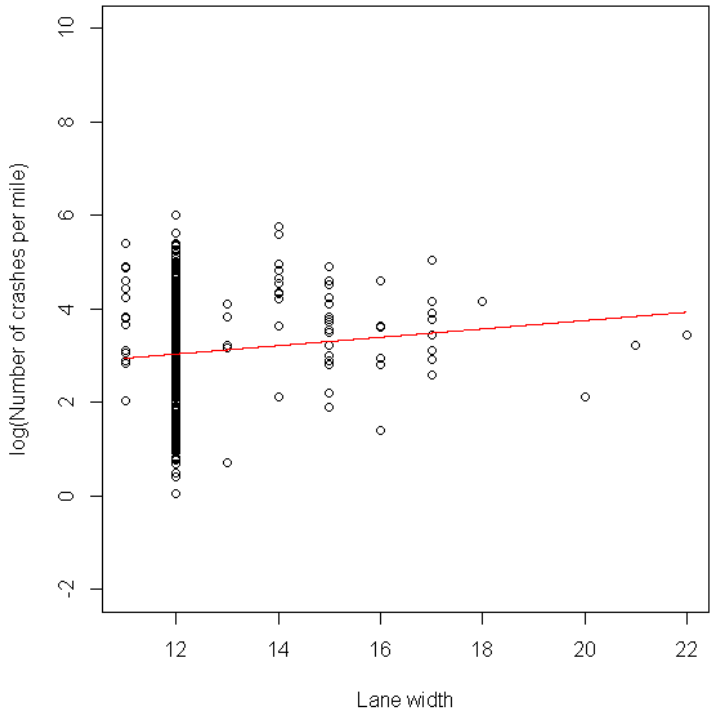
10.4.1 Preliminary Analysis

Figure 10-1 shows the association between explanatory variables and the logarithm of observed crash rates. The generalized linear model (GLM) generally assumes a monotonic relationship between the dependent variable, average crash rate, and explanatory variables through a linearizing link function, $\log(\mu_i)$. As shown in Figure 10-1, the logarithm of the observed crash rate can generally increase or decrease consistently when explanatory variables change as a

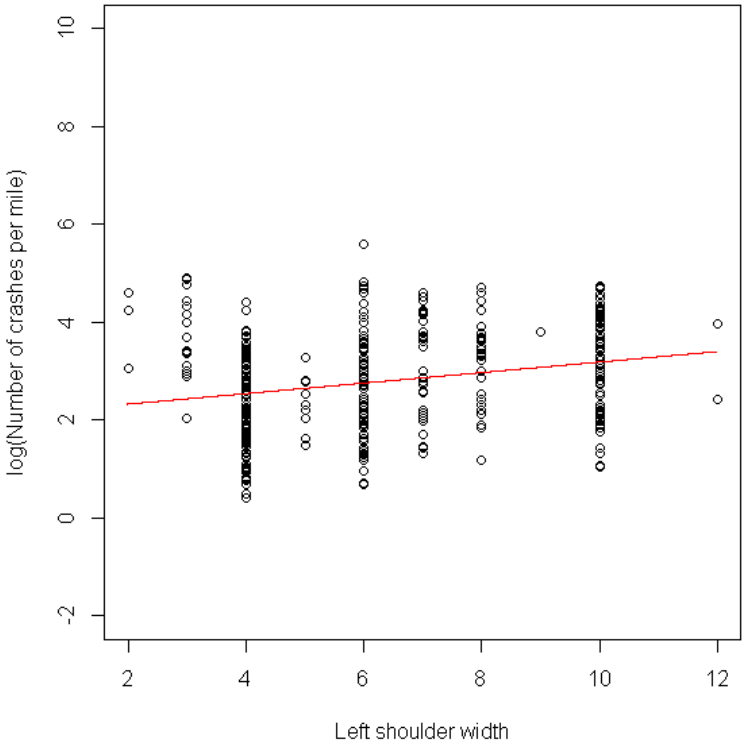
linear function. Therefore, the generalized linear model is considered in the modeling estimation section.



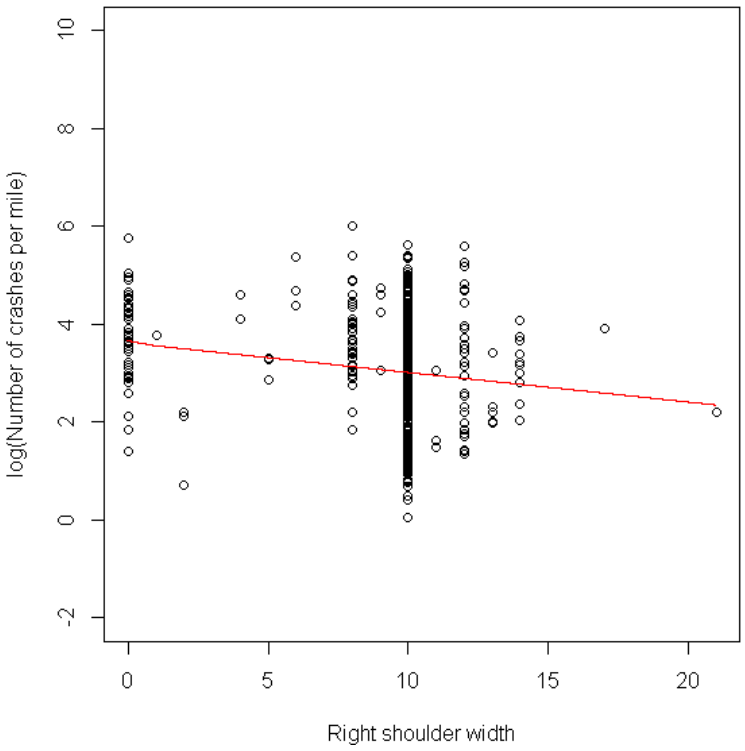
(a)


log(Number of crashes per mile)
-2
0
2
4
6
8
10
12
14
16
18
20
22
Lane width

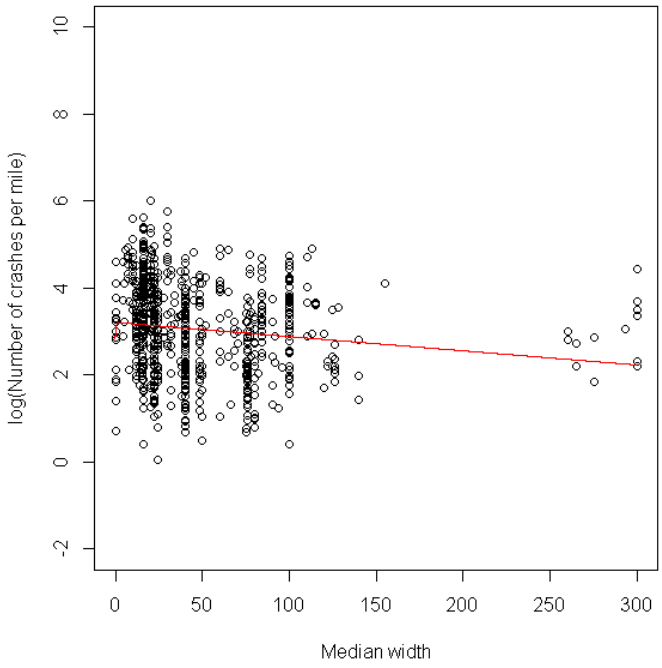
(b)


log(Number of crashes per mile)
-2
0
2
4
6
8
10
2
4
6
8
10
12
Left shoulder width

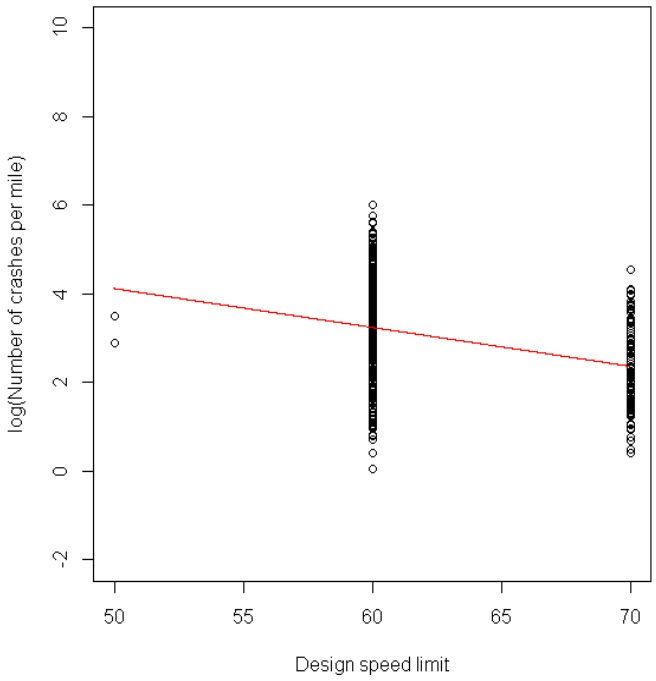
(c)


log(Number of crashes per mile)
-2
0
2
4
6
8
10
0
5
10
15
20
Right shoulder width

(d)


log(Number of crashes per mile)
-2
0
2
4
6
8
10
0
50
100
150
200
250
300
Median width

(e)



(f)

Figure 0-1 Association between explanatory variables and the logarithm of observed crash rates (a) Number of lanes; (b) Lane width; (c) Left shoulder width; (d) Right shoulder width; (e) Median width; (f) Design speed limit.

10.4.2 Modeling Results

This section describes the modeling results for the GLM NB models. For the NB model, a linear mean function of the form shown in Eq. (10-5) is adopted in the modeling process.

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\beta_2 * NL_i + \beta_3 * LW_i + \beta_4 * LSW_i + \beta_5 * RSW_i + \beta_6 * MW_i + \beta_7 * SPD_i} \quad (10\text{-}5)$$

where,

μ_i = estimated number of crashes at segment i over the study period;

L_i = segment length in miles for segment i;

F_i = traffic flow (average daily traffic over the study period) traveling on segment i;

NL_i = Number of lanes for segment i;

LW_i = Lane width in feet for segment i;

LSW_i = Left shoulder width in feet for segment i;

RSW_i = Right shoulder width in feet for segment i;

MW_i = Median width in feet for segment i;

SPD_i =Design speed limit for segment i; and,

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ = estimated coefficients.

Alternatively, if the basic assumption of a generalized linear relationship between the dependent variable (i.e., crash rate) and independent variables is violated, a non-linear mean function of the form shown in Eq. (10-6) can be considered (Lao et al. 2014).

$$\mu_i = \beta_0 L_i \sum_{j=1}^{n} U_j\left(x_{ij}\right) w_j \tag{10-6}$$

where,

μ_i = estimated number of crashes at segment i over the study period;

L_i = segment length in miles for segment i;

$U_j\left(x_{ij}\right)$ = a nonlinear function using the jth explanatory variable;

w_j = the corresponding weight for U_j;

β_0 = estimated coefficient.

As shown in Table 10-2, most explanatory variables except variable left shoulder width were significant when the total crash frequency was predicted. The estimated coefficient values

of the variables from the GLM NB model demonstrate that, first, an increase in AADT was found to be linked to an increase in crash frequency. Road sections with larger lane width, shoulder width, and median width tended to have fewer crashes. Second, increases in number of lanes and speed limit were found to be associated with a decrease in the crash frequency. Third, the estimated coefficients for the explanatory variables remained relatively consistent over the study period.

Table 0-2 Modeling results of total crash counts using GLM NB models for years 2006-2011

	Year 2011		Year 2010		Year 2009	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-6.840	1.302	-6.626	1.462	-6.225	1.327
Log(AADT)	1.154	0.077	1.180	0.084	1.098	0.073
LW	-0.066*	0.040	-0.075	0.038	-0.011*	0.031
NL	-0.056	0.025	-0.053*	0.028	-0.077	0.025
LSW	-0.006*	0.009	-0.015*	0.010	-0.016*	0.009
RSW	-0.050	0.015	-0.057	0.016	-0.036	0.014
MW	-0.004	0.001	-0.005	0.001	-0.003	0.001
SPD	-0.020	0.009	-0.025	0.011	-0.031	0.010
α	0.382		0.433		0.395	
	AIC	4389.61	AIC	3902.843	AIC	4496.93
	SBC	4432.102	SBC	3944.118	SBC	4539.595
	Year 2008		Year 2007		Year 2006	

	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-8.244	1.285	-7.942	1.281	-7.749	1.298
Log(AADT)	1.177	0.069	1.085	0.068	1.293	0.072
LW	0.014*	0.031	0.038*	0.027	-0.071	0.031
NL	-0.073	0.023	-0.014*	0.022	-0.028*	0.022
LSW	-0.013*	0.009	-0.034	0.009	-0.016*	0.009
RSW	-0.026	0.013	-0.013*	0.013	-0.051	0.014
MW	-0.003	0.001	-0.002	0.001	-0.003	0.001
SPD	-0.020	0.009	-0.019	0.009	-0.032	0.010
α	0.354		0.336		0.356	
	AIC	4483.84	AIC	4580.42	AIC	4656.064
	SBC	4526.462	SBC	4623.095	SBC	4698.697

* Insignificant at 0.05 level of significance.

The modeling results for crashes in normal weather conditions and abnormal weather conditions are provided in tables 10-3 and 1--4, respectively. Some explanatory variables were insignificant in modeling crash frequency. The estimated coefficient values of the variables were generally consistent with the modeling results in Table 10-2. In addition, the estimated coefficients for the explanatory variables also remained relatively consistent over the different years.

Table 0-3 Modeling results of crash counts in normal weather conditions using GLM NB models for years 2006-2011

	Year 2011		Year 2010		Year 2009	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-6.792	1.477	-7.695	1.715	-7.524	1.518
logf	1.134	0.087	1.271	0.099	1.277	0.085
LWth	-0.064*	0.044	-0.089	0.044	-0.074	0.036
NLane	-0.064	0.029	-0.078	0.033	-0.091	0.028
LShdWth	-0.016*	0.011	-0.029	0.012	-0.018*	0.010
RShdWth	-0.057	0.017	-0.056	0.018	-0.061	0.015
MdWidth	-0.004	0.001	-0.004	0.001	-0.003	0.001
SPD	-0.022	0.010	-0.027	0.013	-0.031	0.011
Sigma	0.450		0.516		0.446	
	AIC	3897.569	AIC	3385.408	AIC	4007.7
	BIC	3940.062	BIC	3426.684	BIC	4050.365
	Year 2008		Year 2007		Year 2006	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-11.987	1.512	-9.221	1.493	-8.169	1.481
logf	1.468	0.082	1.263	0.079	1.345	0.082
LWth	0.031*	0.036	0.021*	0.035	-0.067	0.035
NLane	-0.123	0.025	-0.049	0.024	-0.058	0.025

LShdWth	-0.008*	0.010	-0.028	0.010	-0.008*	0.010
RShdWth	-0.031	0.015	-0.014*	0.015	-0.047	0.015
MdWidth	-0.005	0.001	-0.002	0.001	-0.003	0.001
SPD	-0.019*	0.011	-0.033	0.011	-0.041	0.011
Sigma	0.377		0.368		0.407	
	AIC	3812.526	AIC	4009.773	AIC	4150.933
	BIC	3855.148	BIC	4052.449	BIC	4193.565

* Insignificant at 0.05 level of significance.

Table 0-4 Modeling results of crash counts in abnormal weather conditions using GLM NB models for years 2006-2011

	Year 2011		Year 2010		Year 2009	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-10.159	1.876	-8.058	2.040	-6.724	1.868
logf	1.282	0.111	1.141	0.117	0.874	0.102
LWth	-0.063*	0.058	-0.057*	0.051	0.075*	0.041
NLane	-0.045*	0.035	-0.016*	0.037	-0.056*	0.035
LShdWth	0.015*	0.013	0.013*	0.014	-0.012*	0.013
RShdWth	-0.030*	0.021	-0.055	0.021	0.015*	0.019
MdWidth	-0.005	0.001	-0.008	0.002	-0.003	0.001
SPD	-0.014*	0.012	-0.018*	0.015	-0.028	0.013

Sigma	0.475		0.57		0.542	
	AIC	2962.038	AIC	2769.402	AIC	2948.017
	BIC	3004.53	BIC	2810.678	BIC	2990.682
	Year 2008		Year 2007		Year 2006	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-5.540	1.762	-8.441	1.782	-10.234	1.841
logf	0.846	0.094	0.895	0.095	1.294	0.102
LWth	-0.012*	0.043	0.041*	0.037	-0.064*	0.044
NLane	-0.004*	0.031	0.036*	0.030	0.014*	0.030
LShdWth	-0.018*	0.012	-0.045	0.012	-0.029	0.012
RShdWth	-0.011*	0.019	-0.011*	0.018	-0.050	0.019
MdWidth	-0.002	0.001	-0.002	0.001	-0.004	0.001
SPD	-0.022*	0.013	0.002*	0.013	-0.015*	0.014
Sigma	0.504		0.479		0.476	
	AIC	3275.947	AIC	3230.351	AIC	3206.355
	BIC	3318.569	BIC	3273.027	BIC	3248.988

* Insignificant at 0.05 level of significance.

10.5 DRIVE Net Application

In the DRIVE Net system, the safety performance analysis function is under the "Safety Performance" tab. The panel is named "Safety Performance Regression" (see Figure 10-2).

Three sub-functions are implemented in this panel: Incident Frequency, Estimated Crash Mean, and Potential Safety Improvement Index.

As stated previously, within a selected time range and corridor, incident frequency shows the average accident rate per mile per year. The accident/incident data were from WSDOT's WITS database. The incident frequency level ranges from Level A to Level F, where Level A (light green) corresponds to the least incidents and Level F (dark red) corresponds to the most incidents observed historically. Figure 10-3 shows an example of this function.

The estimated crash mean function implements the safety performance estimation by using the roadway characteristics described above. This function has nothing to do with the time range. It uses the values in the eight textboxes that indicate the weights of the characteristics. The default values are calculated in the regression modeling. However, these values can be defined by users. In other words, users can implement their own model to evaluate the safety performance of Washington state corridors. Like the Incident Frequency function, this function also uses Level A to Level F to represent least safe to most safe. Figure 10-4 shows an example of this function.

The Potential Safety Improvement Index implements the EB method for modeling. In this function, both the historical incident data and the characteristics of the selected corridor are input. The output format also uses the six different colors representing Level A to Level F to show the potential safety improvement index on the map, where Level A indicates the segment with the least potential for improvement, and Level F indicates the segment with the most potential for improvement. Figure 10-5 shows an example of this function.

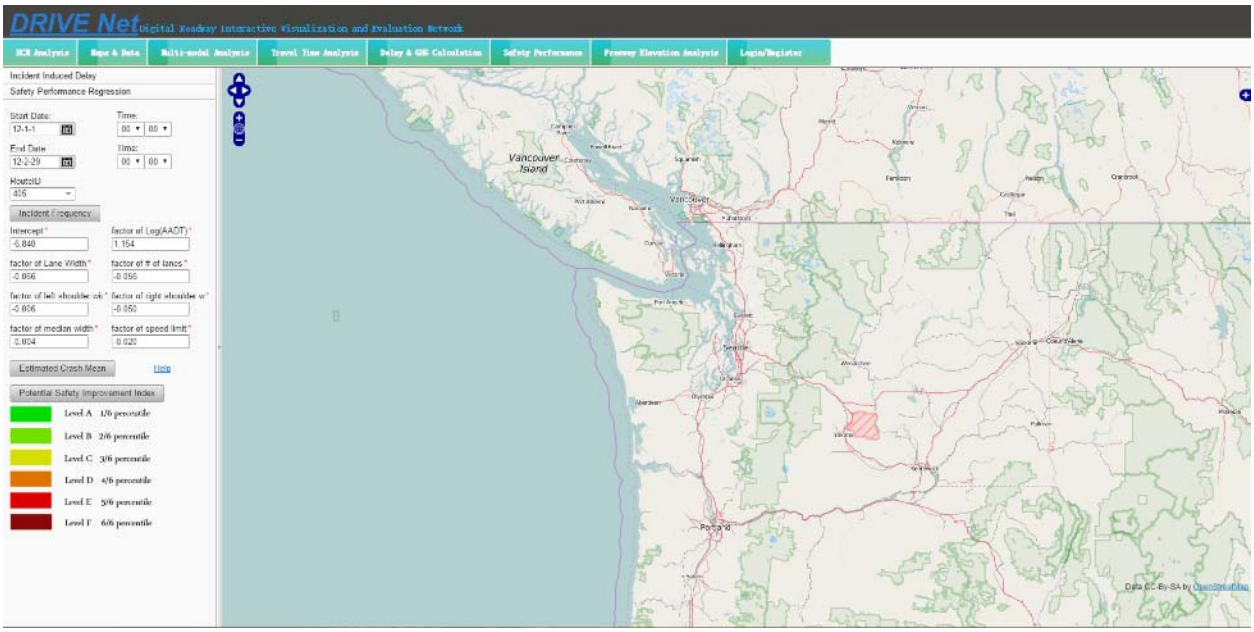


Figure 0-2 Interface design of the implementation of Safety Performance

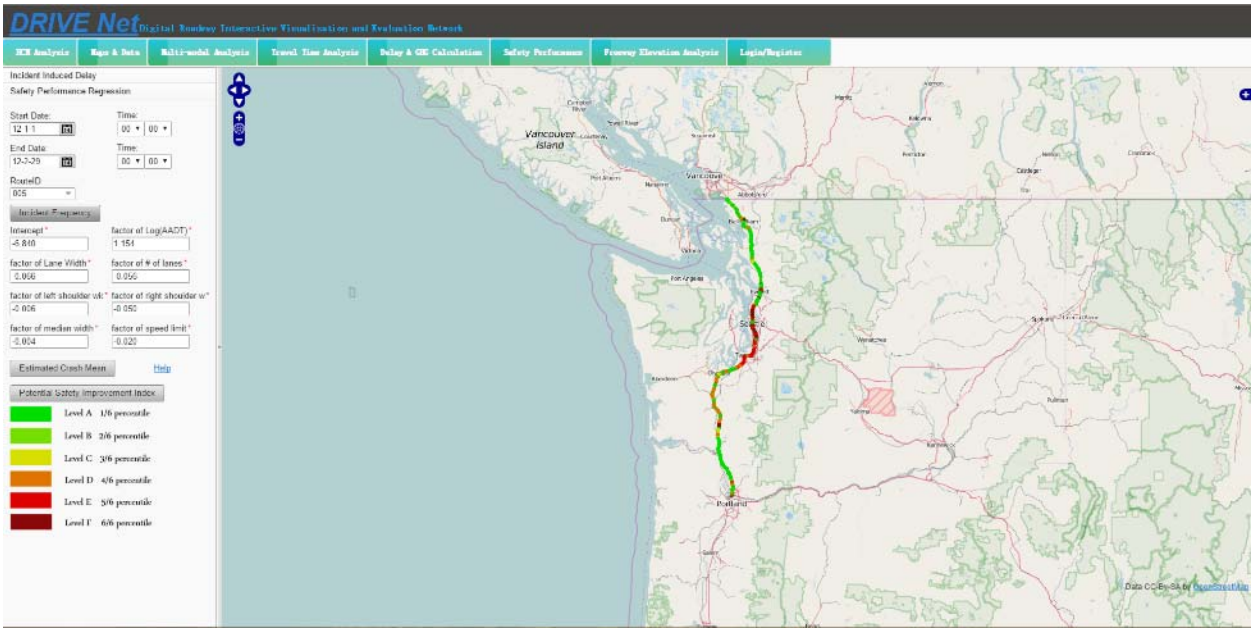


Figure 0-3 An example of the Incident Frequency function

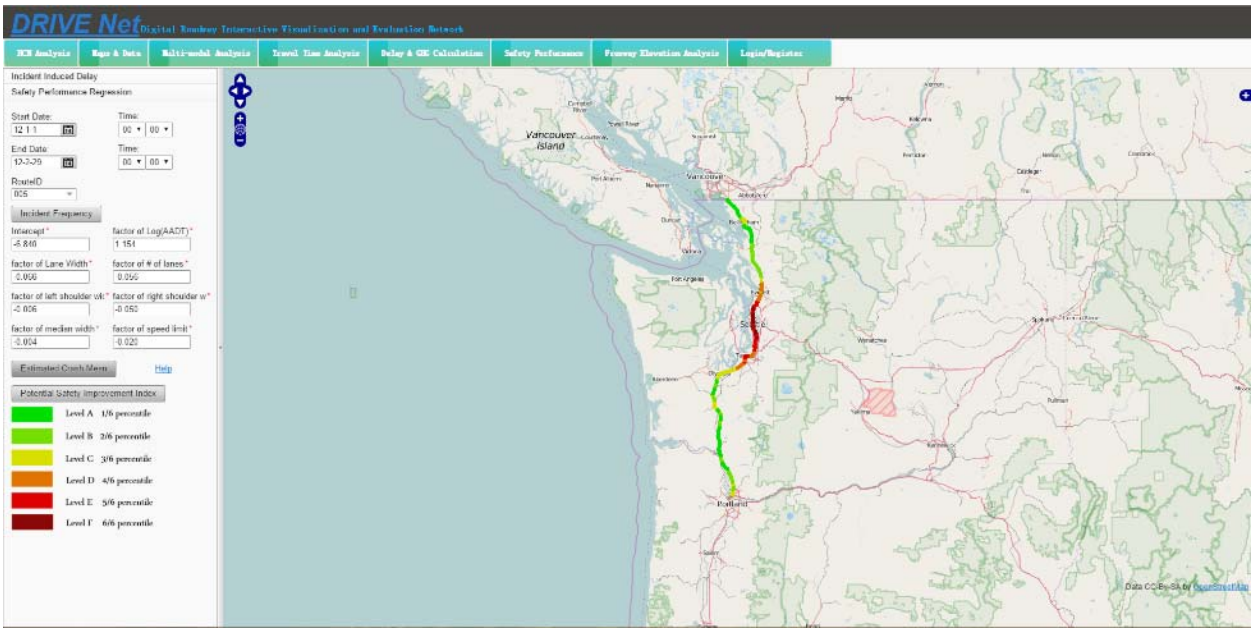


Figure 0-4 An example of the Estimated Crash Mean function.

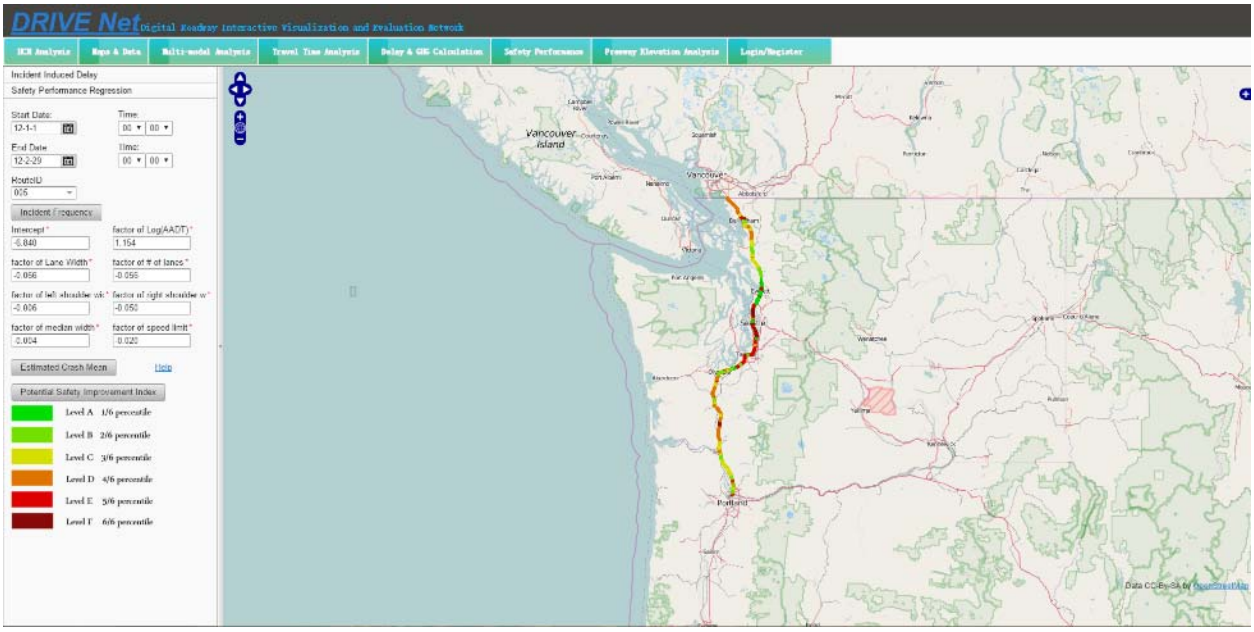


Figure 0-5 An example of the Potential Safety Improvement Index function

Chapter 11 Data Sharing Function

To conduct some in-depth DRIVE Net analysis, some raw data are available for downloading. Three kinds of data sets can be downloaded: INRIX raw data, HERE raw data, and freeway loop raw data. The downloading process can be divided into three sub-processes: login to the system, choose desired corridors and a time range, and email the raw data. The download process for three data sources is described separately.

11.1 INRIX Raw Data Downloader

1. Get online

Find out how this website works and how to access the info when on a public network. All functions of the DRIVE Net system are open access for all users except the data download segments.

2. Set up a DRIVE Net account

In order to download INRIX data, users must first register with the DRIVE Net system. To facilitate and simplify the data download process, authorized login user name and password have been fixed as DRIVE Netdemo@gmail.com and DRIVE Net2014, respectively.

3. Under the Maps & Data tab, choose the INRIX Raw Data Downloader.

Figure 11-1 is the interface for downloading INRIX data, and Figure 11-2 shows an overall view of the INRIX data downloader.

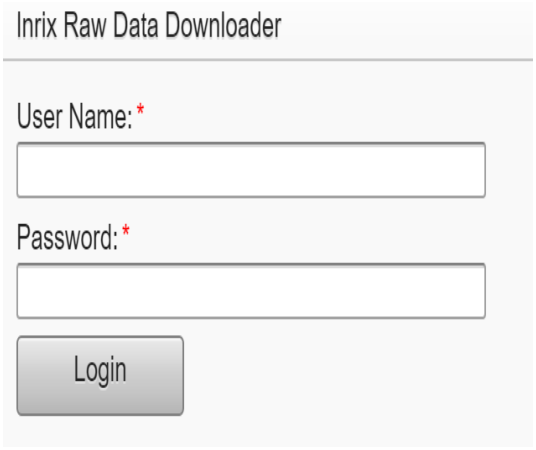


Figure 0-1 User interface for the INRIX Raw Data Downloader

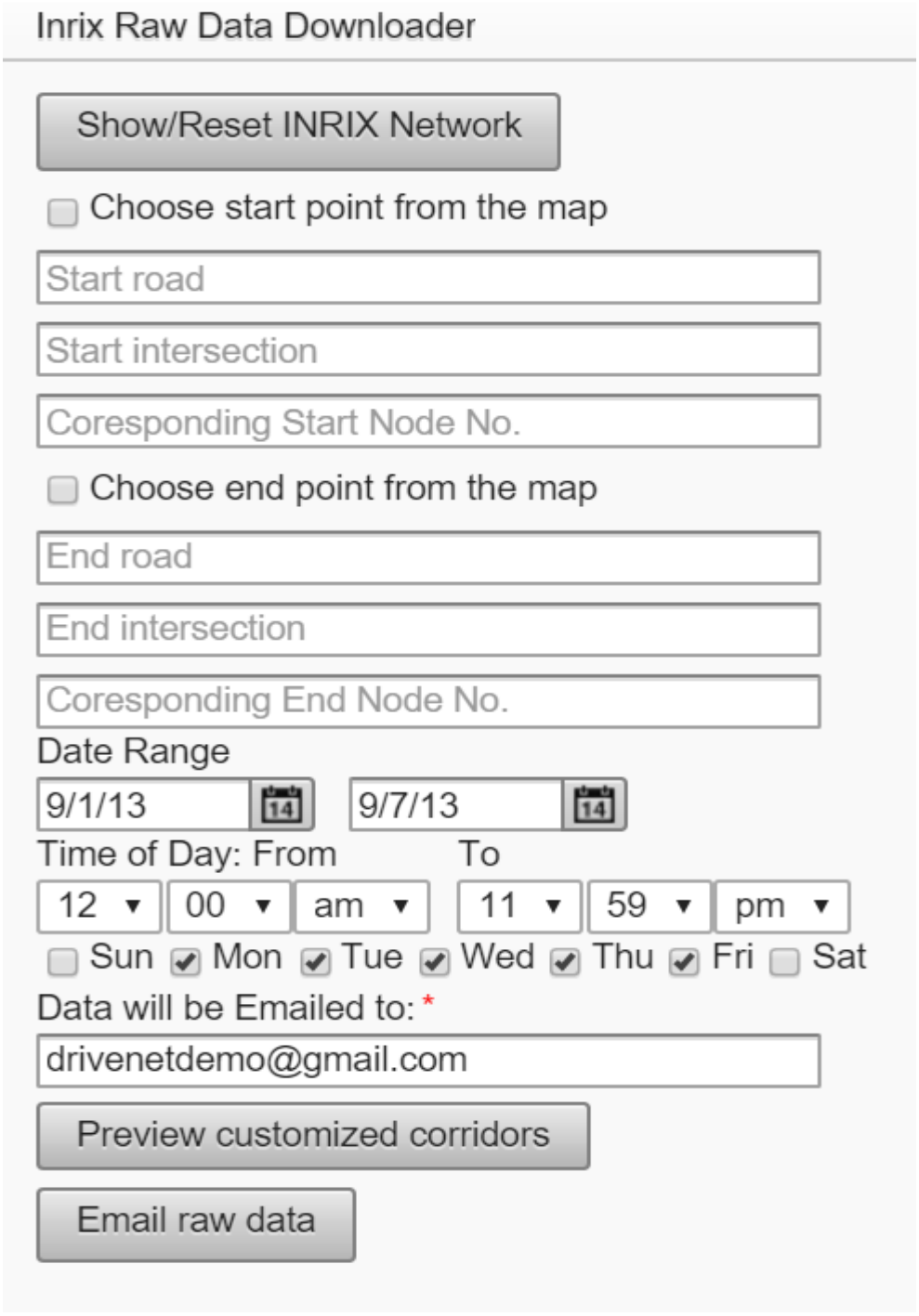
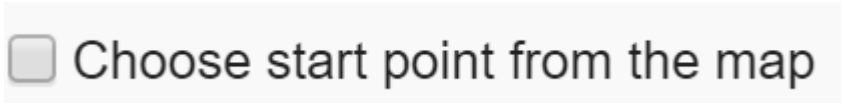


Figure 0-2 Overall view of INRIX Data Downloader

4. Download INRIX data (scenario case):

4a) 

To choose a starting point, input road information manually or select points from the map. For simplicity, choosing points by clicking on the map is preferred.

Now, click the check box:

To click the street on the map itself, users need prior knowledge about street location, and this should only be applied when the user is sure that s/he is in the right area of analysis. Note that the search bar does not give options like a GPS car system; users need to know where they want to analyze. The crucial components are the start road, start intersection, and corresponding start node number. Other than knowing the location on the map and clicking on it, users can manually define it by plugging in the start road address, start intersection, and corresponding start node number. Corresponding text boxes are shown in Figure 11-3.

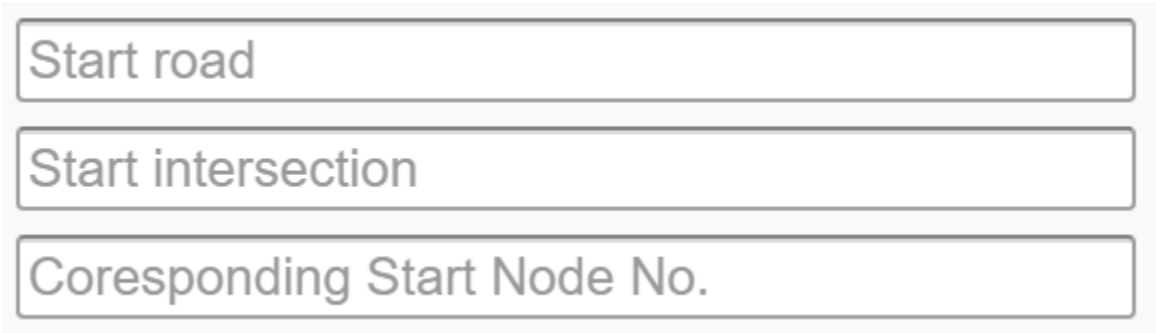


Figure 0-3 Textboxes for selecting roads for the INRIX Downloader

4b) For the sake of simplicity, we randomly clicked start and end points on the INRIX network map, shown in Figure 11-4. Figure 11-5 shows the chosen corridor, and Figure 11-6 shows corresponding information for the selected corridor. To download corridor data, the time scope must also be specified. The default time scope is shown in Figure 11-7.

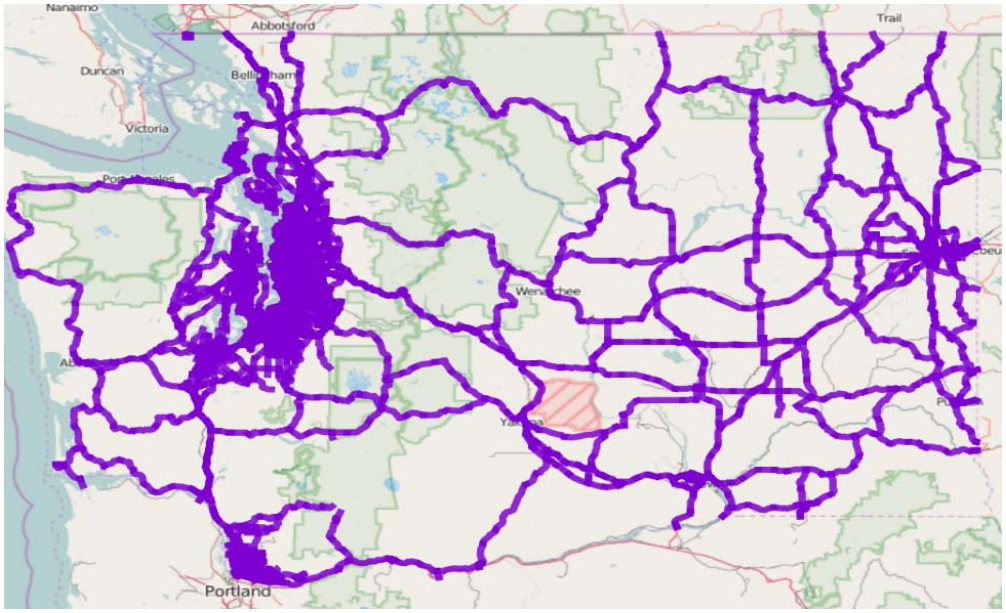


Figure 0-4 INRIX network map

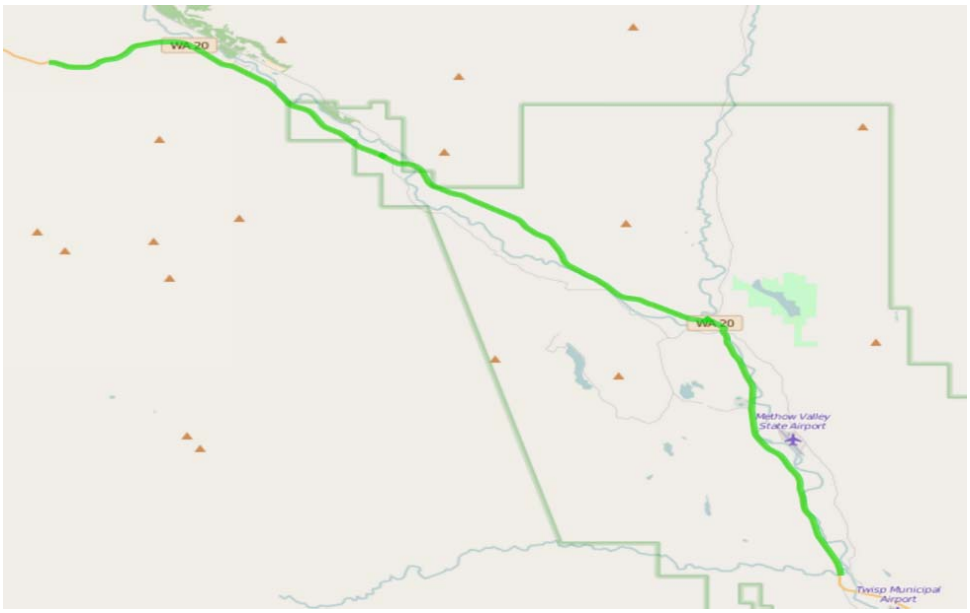


Figure 0-5 Randomly selected corridor from the INRIX network map

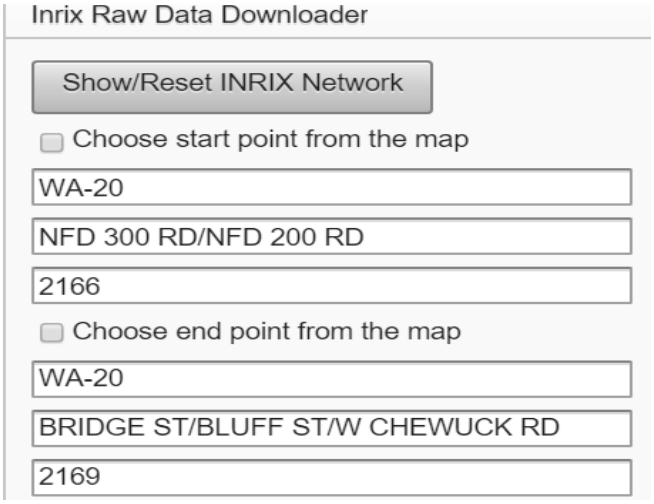


Figure 0-6 Corridor information corresponding to Figure 11-5

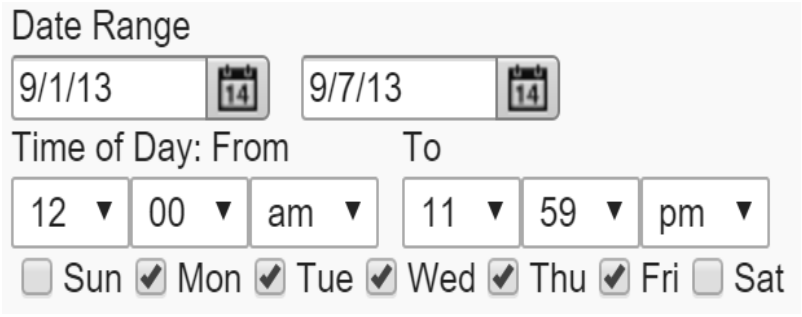


Figure 0-7 Time selection widget

4c) The next step is to download the data. The data downloader will send two messages to the email registered with the DRIVE net system. The first one will contain a data-download request confirmation, and the second one will contain the download hyperlink. Table 11-1 is a data segment from an INRIX raw data set. It shows that time, Traffic Message Channel (TMC), and speed have been archived in the INRIX data set. The time column in Table 11-1 represents the time range shown in Figure 11-7. TMC displays corridor location, while average speed for the given corridor and time are listed in the third column.

Table 0-1 Subset of downloaded INRIX raw data

Time	TMC	Speed
41513.80972	114+11140	31
41514.80972	114+04325	33
41515.80972	114+04597	45
41516.80972	114+05450	51

11.2 HERE Raw Data Downloader

1. Get online

Find out how this website works and how to access the info when on a public network. All the functions of DRIVE Net system are open access for all users except data download segments.

2. Set up a DRIVE Net account

To download HERE data, the first step is to register with the DRIVE Net system. To facilitate and simplify the data download process, the authorized login user name and password have been fixed as DRIVE Netdemo@gmail.com and DRIVE Net2014, respectively.

3. Under the Maps & Data tab, choose the HERE Raw Data Downloader.

Figure 11-8 is the interface for downloading HERE data, and Figure 11-9 shows the overall view of the HERE data downloader. As the HERE data set records vehicles' travel times, values vary with vehicle categories. The default input for the HERE Data Downloader is average travel time with all vehicles, shown with a red rectangle in Figure 11-9.

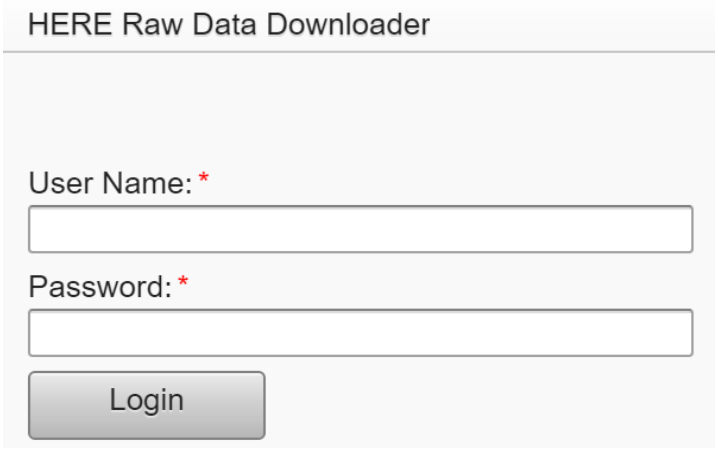
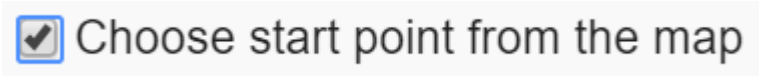


Figure 0-8 User interface for HERE Raw Data Downloader

4. Download HERE data in this application (scenario case):

4a) 

To choose a starting point, input road information manually or select a point from the map. For simplicity, choosing a point by clicking on the map is preferred.

Now, click the check box: 

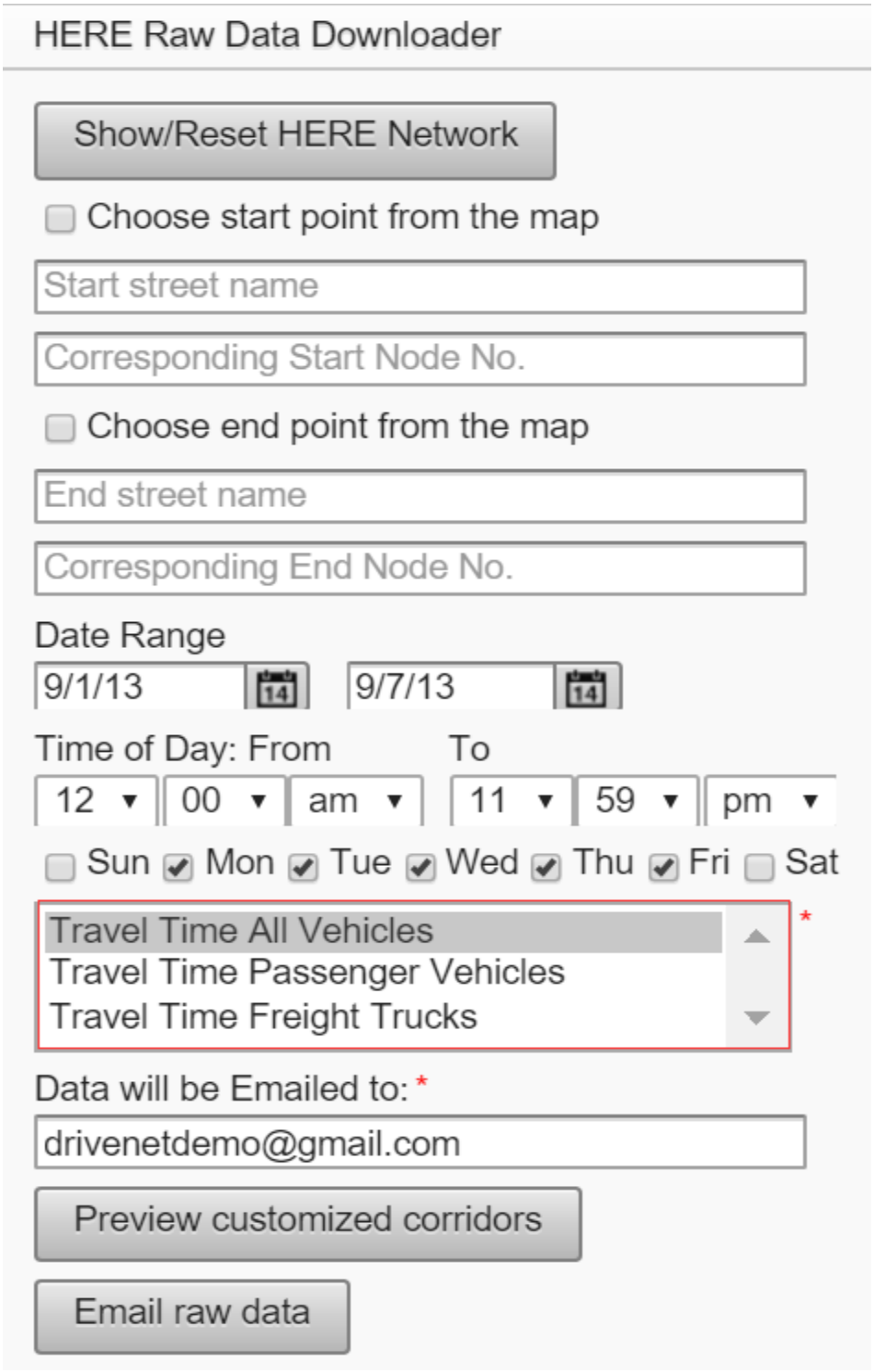


Figure 0-9 Overall view of the HERE Data Downloader

Clicking the street on the map requires users to know which street and direction they want. This method for getting a street location can only be applied when the user is sure that s/he is in the right area for analysis. Note that the search bar does not give options like a GPS car system; users need to know where they want to analyze. The crucial components are the start road, start intersection, and corresponding start node number. Users can also enter a location manually by entering the start road address, start intersection, and corresponding start node number. Corresponding text boxes are shown in Figure 11-10.

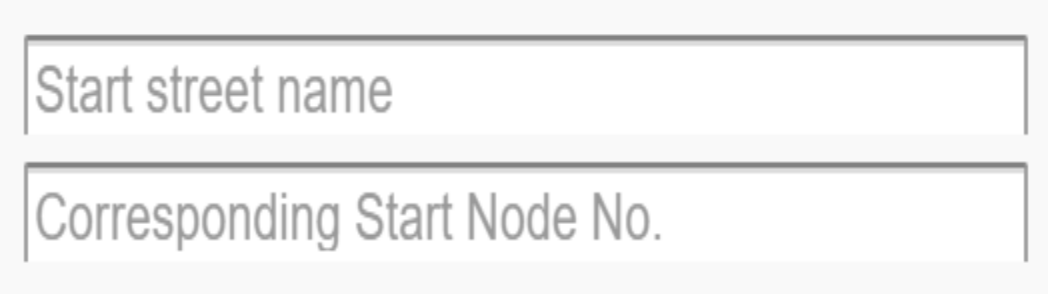


Figure 0-10 Text boxes for selecting a road for the HERE Downloader

4b) For the sake of simplicity, we have randomly clicked the start and end points on the HERE network map, shown in Figure 11-11. Figure 11-12 shows the chosen corridor and Figure 11-13 shows the corresponding information for the selected corridor. To download corridor data, the time scope must also be specified. A default time scope is shown in Figure 11-14.

4c) The next step is to download the data. The data downloader sends two messages to the registered email address. The first one contains a data-download request confirmation. The second one contains the download hyperlink. Table 11-2 shows a segment from HERE raw data sets. The time column in Table 11-2 represents the time range entered the Figure 11-14.

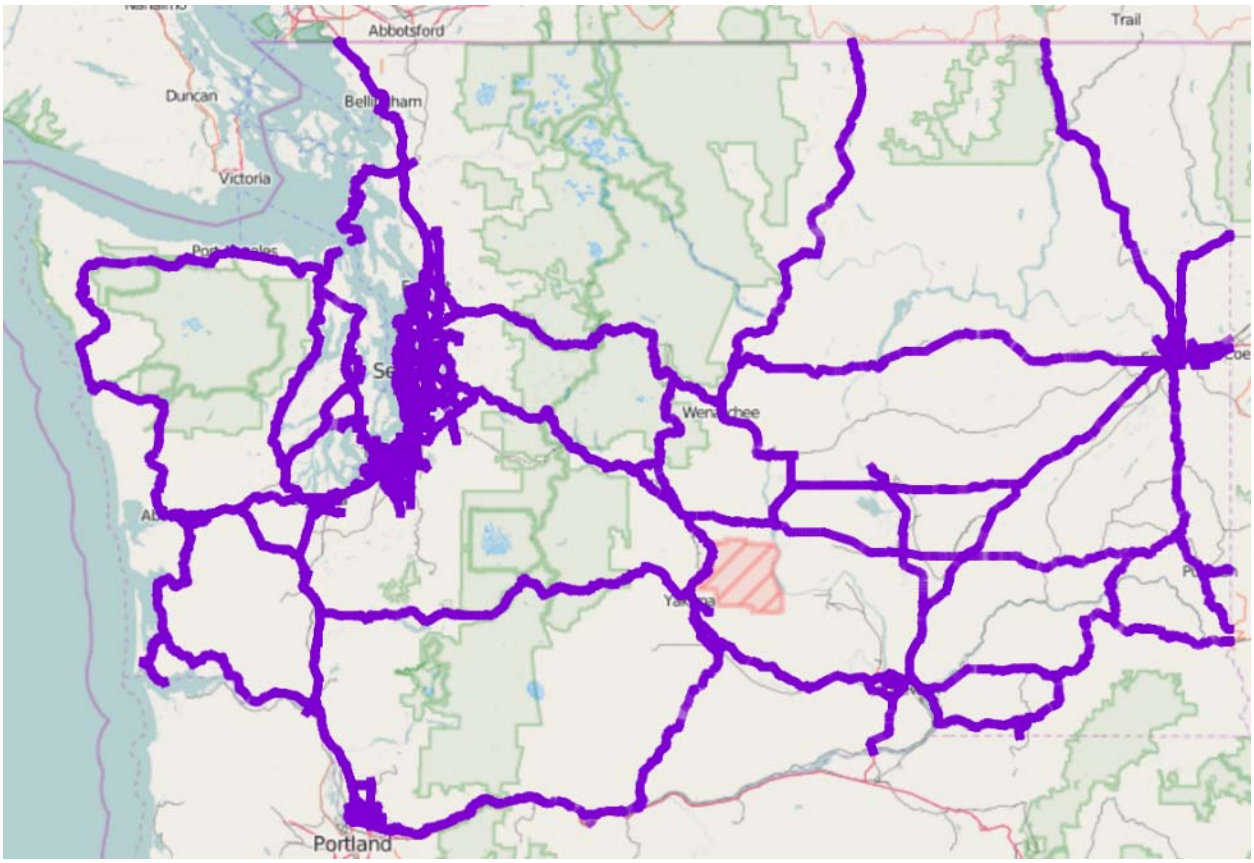


Figure 0-11 HERE network map

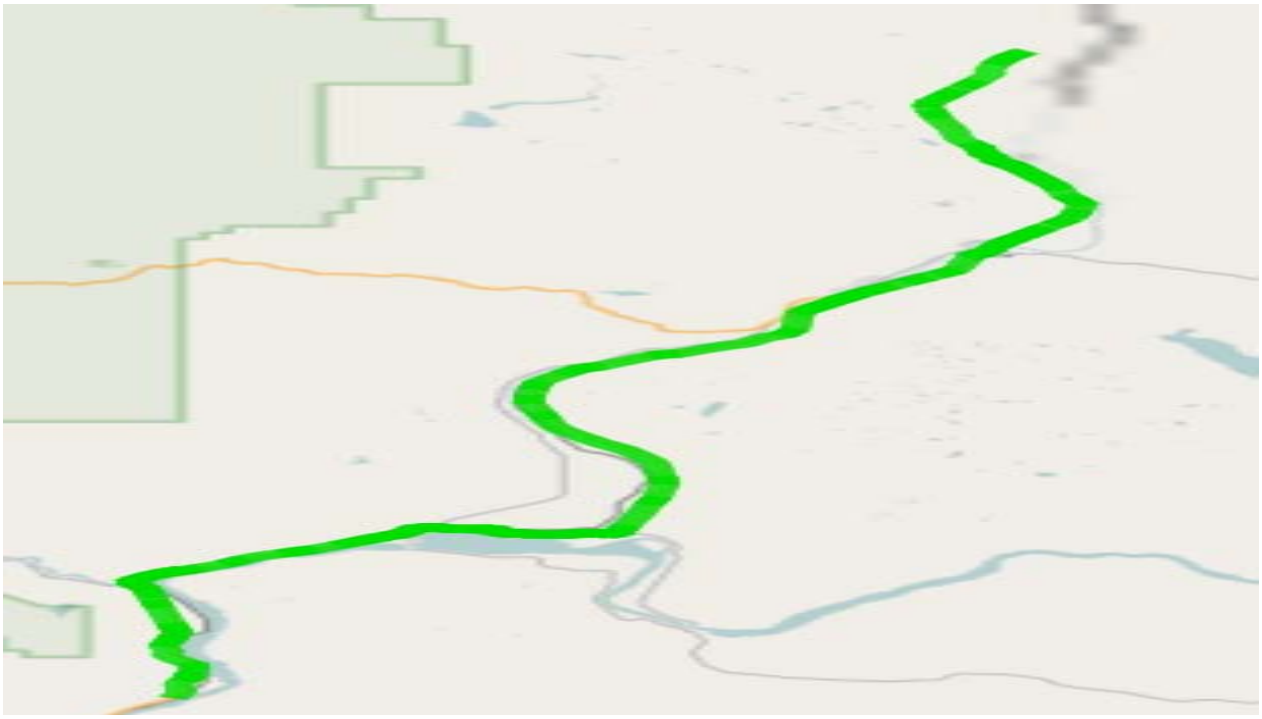


Figure 0-12 Randomly selected corridor from HERE network map

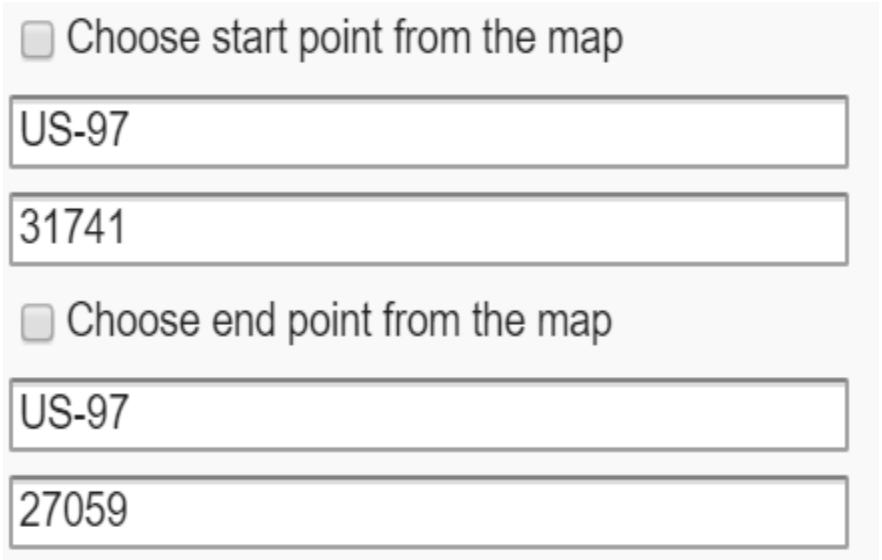


Figure 0-13 Corridor information corresponding to Figure 11-12

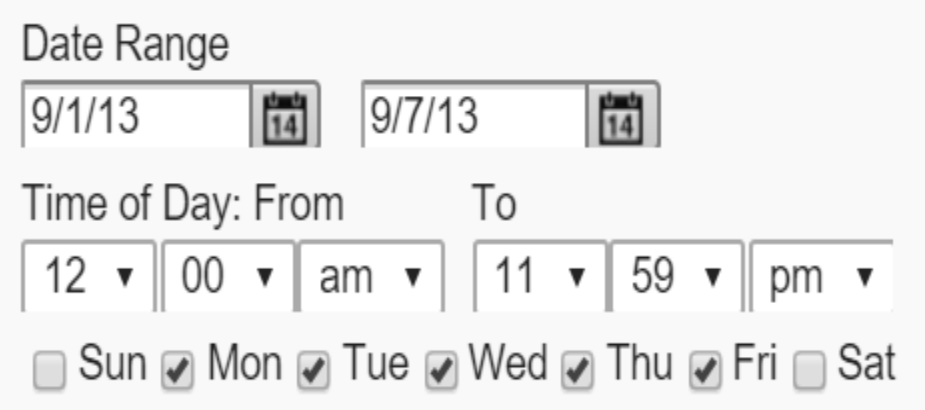


Figure 0-14 Time selection widget

Table 0-2 Subset of downloaded HERE raw data

TMC	DateTimeStamp	TravelTimeAllVehicles	TravelTimePassengerVehicles	TravelTimeFreightTrucks
114N10346	09/01/13 16:35	27	27	0
114N10753	09/02/13 16:35	6	6	0
114N10346	09/03/13 16:35	27	27	0
114N10753	09/04/13 16:35	6	6	0

11.3 Loop Raw Data Downloader

1. Get online

Find out how this website works and how to access the info when on a public network. All the functions of DRIVE Net system are open access for all users except data download segments.

2. Set up a DRIVE Net account

3. Under the Maps & Data tab, choose the Loop Raw Data Downloader.

Figure 11-15 shows the interface for downloading Loop data, and Figure 11-16 shows an overall view of the Loop Data Downloader. Because the Loop dataset records vehicles' travel times, values vary with vehicle categories.

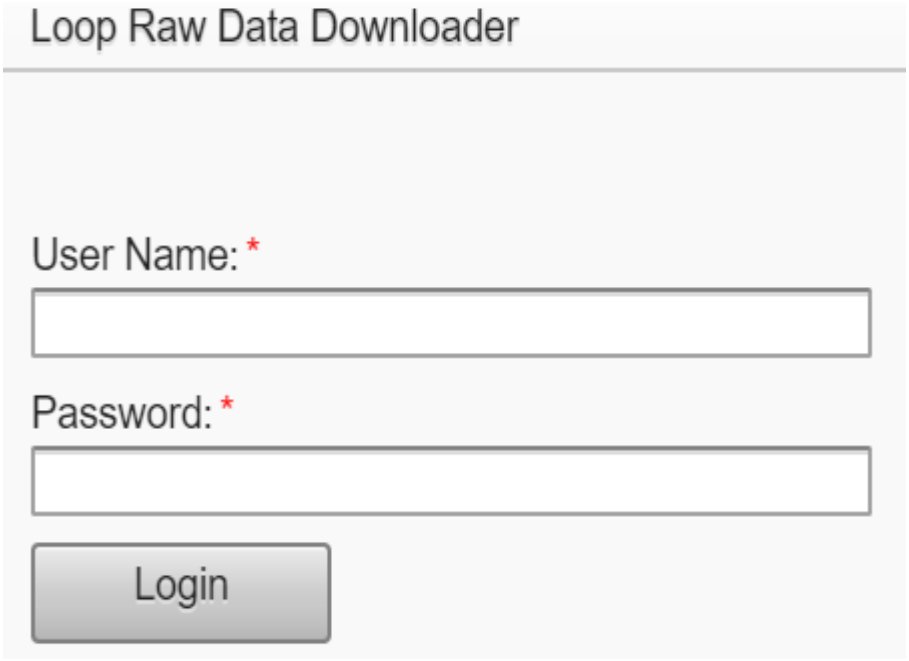
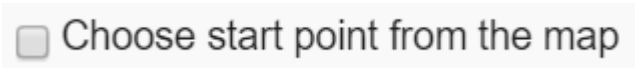
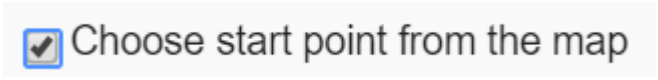


Figure 0-15 User interface for the Loop Raw Data Downloader

4. Download loop data in this application (scenario case):

4a) 

To choose starting point, users must input road information manually or select points from the map. For simplicity, choosing points by clicking on the map is preferred.

Now, click the check box: 

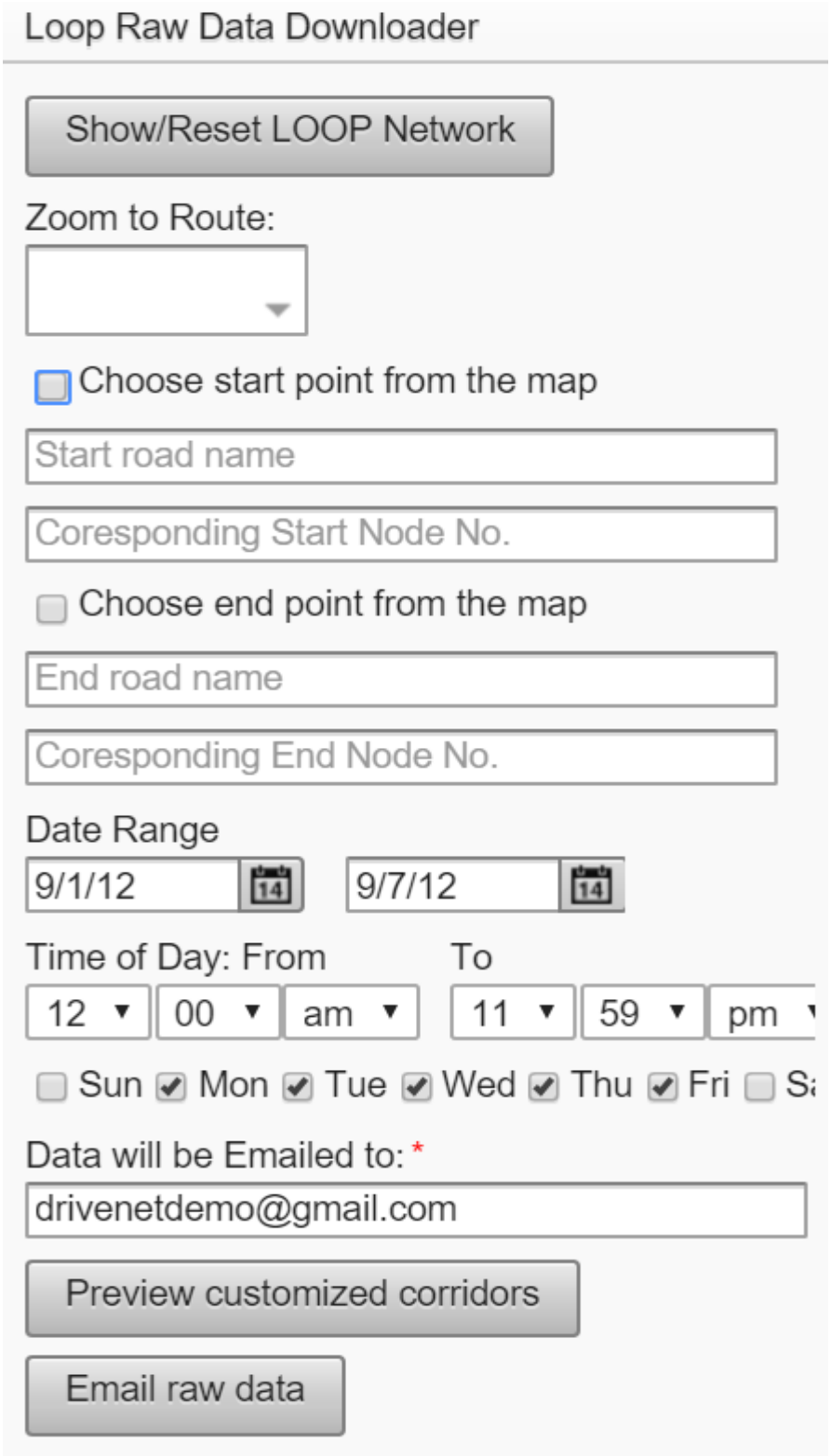


Figure 0-16 Overall view of the Loop Data Downloader

To click streets on the map itself, users need prior knowledge about the street and direction they intend. This can only be applied when the user is sure that s/he is in the right area for analysis. Note that the search bar does not give options like a GPS car system; users need to know where they want to analyze. The crucial components are the start road, start intersection, and corresponding start node number. Other than knowing the location on the map and clicking on it, users can also manually enter the start road address, start intersection, and corresponding start node number. Corresponding text boxes are shown in Figure 11-17.

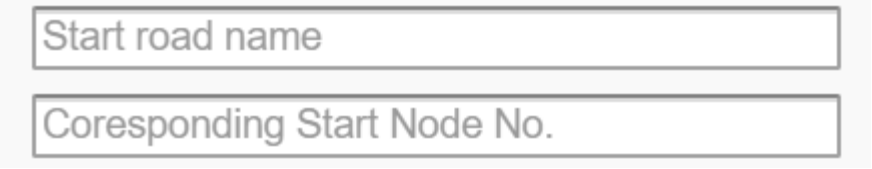


Figure 0-17 Text boxes for selecting a road for the Loop Downloader

4b) For the sake of simplicity, we randomly clicked start and end points on the Loop network map, shown in Figure 11-18. Figure 11-19 shows the chosen corridor, and Figure 11-20 shows corresponding information for the selected corridor. To download corridor data, the time scope must also be specified. The default time scope is shown in Figure 11-21.

4c) The next step is to download data. The data downloader sends two messages to the registered email address. The first one contains a data-download request confirmation, and second one contains the download hyperlink. Table 11-3 shows data from a segment of a raw loop data set. The time column in Table 11-3 represents the time range entered Figure 11-21.

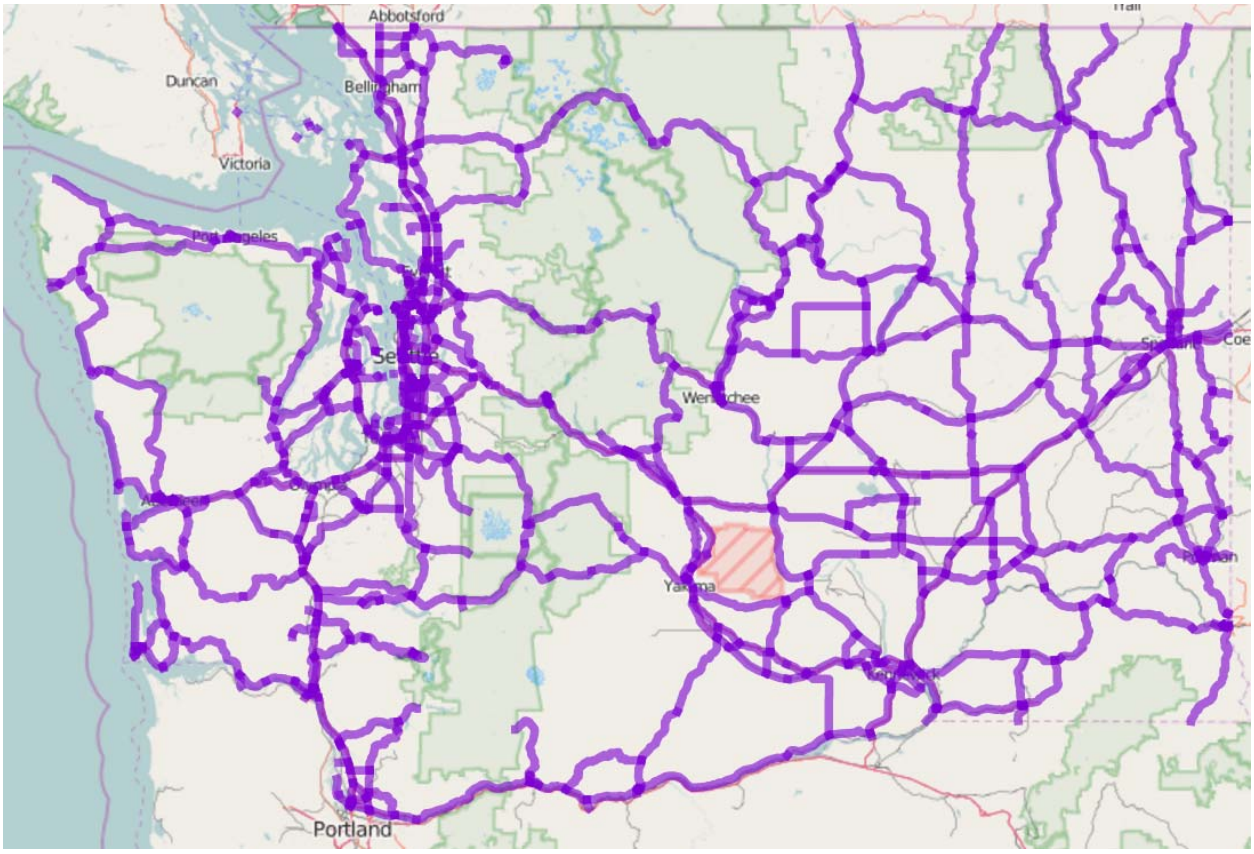


Figure 0-18 Loop network map

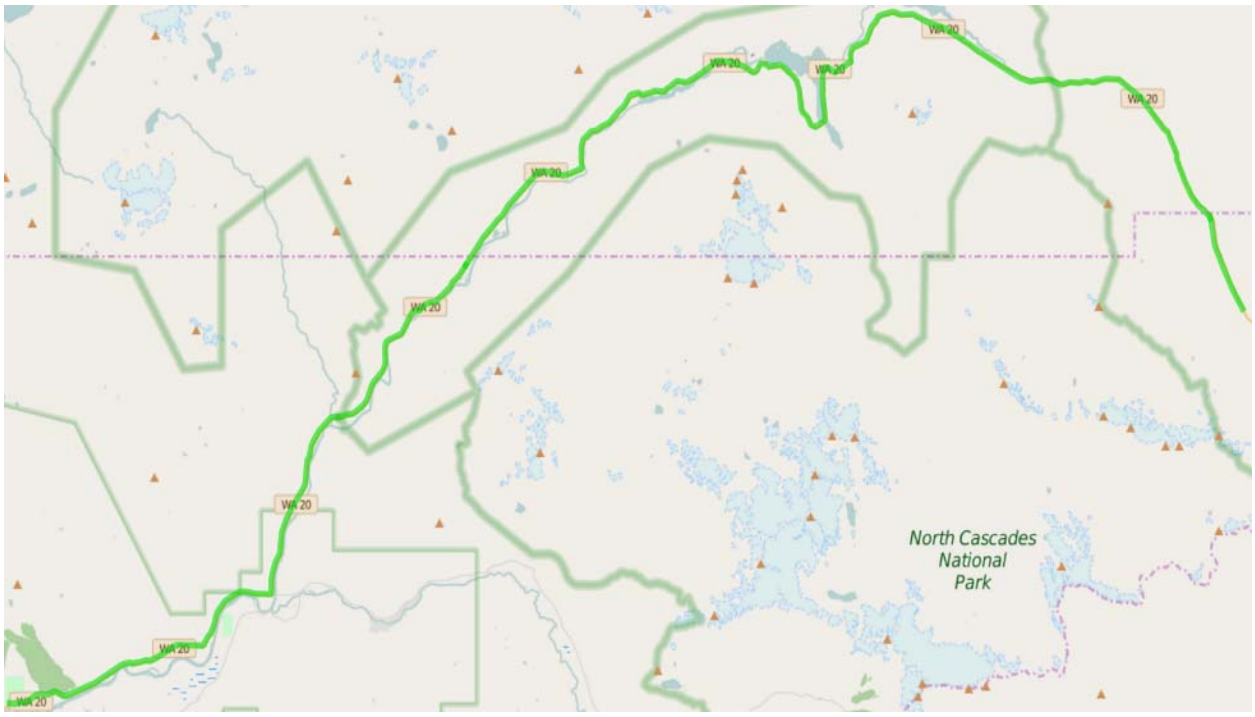


Figure 0-19 Randomly selected corridor from the Loop network map

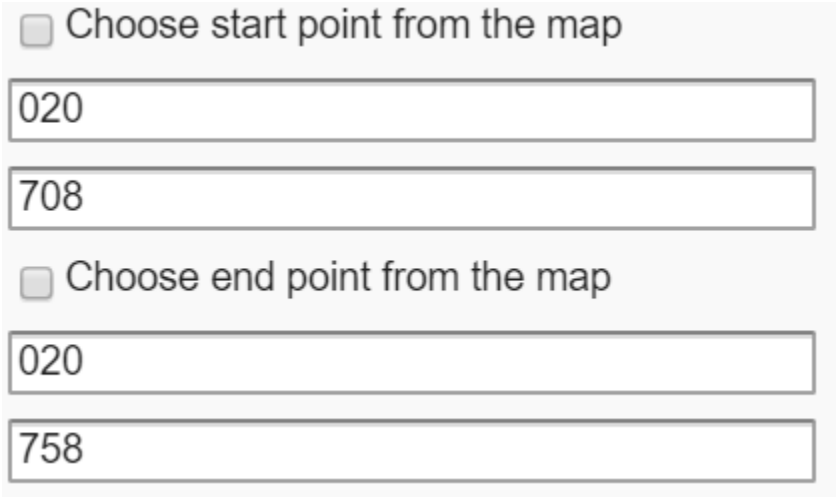


Figure 0-20 Corridor information corresponding to Figure 11-19

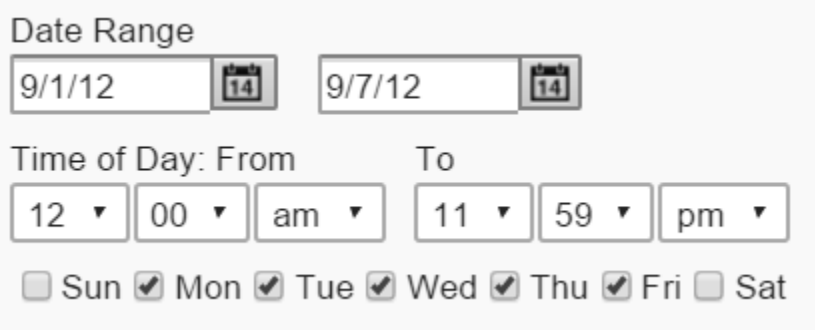


Figure 0-21 Time selection widget

LoopID	DateTimeStamp	Incident	Data	Flag	Volume	Scan
4451	09/03/12 00:00	0	1	6	0	0
4451	09/04/12 00:00	0	1	6	0	0
4451	09/05/12 00:00	0	1	6	0	0
4451	09/06/12 00:00	0	1	6	0	0

Figure 0-22 Subset of downloaded Loop raw data

Chapter 12 Conclusions and Recommendations

12.1 Conclusions

The push toward smarter transportation planning, management, and decision making has increased significantly in recent years. Although the quantity and variety of data available to transportation engineers and decision makers is rapidly increasing, there is a significant need for data management, analysis, and visualization tools to make those data accessible for decision support. The challenges associated with managing and integrating numerous large and heterogeneous data sets are such that much of the internal and third party data now available to WSDOT and other state transportation agencies are not being used to the agencies' full benefit. Consequently, WSDOT has a significant need for a tool that has the data storage and analysis capability to allow fast analysis of data from multiple sources in support of project planning, scoping, design, construction, performance analysis, reporting, and system maintenance activities.

DRIVE Net, a powerful, web-based transportation analytics platform, has been developed by the UW STAR Lab to address the aforementioned challenges. DRIVE Net is a data management, quality control, analysis, and visualization platform that has the ability to layer a diverse spectrum of spatial and temporal data sets on an online digital roadway map. This second version of the platform offers the ability to handle more complex computational tasks, perform large-scale spatial processing, and support data sharing services. The result is a more stable and interoperable system for processing, analyzing, visualizing, and sharing a diverse collection of transportation data that previously did not exist. DRIVE Net is designed to support the current and future analytical needs of WSDOT, and it provides a powerful and up-to-date alternative to WSDOT's current traffic data analysis tools.

A number of enhancements and new features have been added to the current version of DRIVE Net. Specifically, a variety of new data sets have been added to the system, including car sharing data, ferry traffic, public transit, and pedestrian traffic data. Additionally, new analytical functions have been developed to increase the safety and performance analysis functionality of DRIVE Net. Data sharing capabilities have been enhanced to provide easy access to a number of

preprocessed data resources, as well as role-based access control. In summary, the updated version of DRIVE Net offers improved performance and reliability, additional and enhanced functionality, and access to a number of new and expanded data resources. This system will support WSDOT's mission in the future by addressing many of the challenges associated with processing, managing, and integrating large and heterogeneous transportation data sets and making those data accessible for decision support.

12.2 Recommendations

To facilitate future research, the following recommendations are made:

(1) As the capabilities and user base of DRIVE Net expands, managing the hardware and software in-house and ensuring a high degree of availability will become increasingly challenging and time consuming. Furthermore, the size and coverage of many new and emerging data sources applicable to transportation, as well as the high rate of data ingestion required, will present an increasing challenge in terms of computational load. Because of this, WSDOT may look to cloud computing services in the future to alleviate the computational burden and achieve real-time analytics.

(2) Again, because of the scope and magnitude of new and emerging data sources, distributed computing and data management software should be considered. For example, Apache Hadoop is a distributed and scalable programing framework designed to work with very large data sets and support high data ingestion rates, powerful analytical tools, and a high degree of availability. The application of such technology will likely improve system performance significantly.

(1) Although DRIVE Net has addressed many of the challenges associated with large-scale transportation data integration and analysis, one key challenge remains. Specifically, there is a need for enhanced capacity to handle heterogeneous data sources in a fully unified way. Future work will address this issue by creating an advanced data schema for the many current and emerging transportation data sources.

References

Abbess, C., Jarrett, D. & Wright, C. C., Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the "Regression-To-Mean" Effect. Traffic Engineering & Control, 22, 535-542, 1981.

Al-Deek, H. M., Venkata, C., and Chandra, S. R. New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *In Transportation Research Record: Journal of the Transportation Research Board*, No. 1867- 1, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 116-126.

Alta Planning and Design. "National Bicycle and Pedestrian Document Project." Alta Planning and Design, Inc. 2006. www.altaplanning.com/. Accessed January 31, 2011.

Bargar, A., Gupta, A., Gupta, S., and Ma, D. Interactive Visual Analytics for Multi-City Bike share Data Analysis. In The 3rd International Workshop on Urban Computing (UrbComp 2014), New York, USA, 2014.

Barry M. and Card B. Boston's Massachusetts Bay Transit Authority (MBTA). [Online]. Available: http://mbtaviz.github.io/, on April 5, 2016.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Buuren, S., Oudshoom, K. Flexible multivariate imputation by MICE. TNO Prevention and Health. 1999.

CAO, X., XU, Z. & HUANG, A. Y. 2012. Safety Benefits of Converting HOV Lanes to HOT Lanes: Case Study of the I-394 MnPASS. Institute of Transportation Engineers. ITE Journal, 82, 32-37.

Chang, H., Park, D., Lee, Y., and Yoon, B. Multiple time period imputation technique for multiple missing traffic variables: nonparametric regression approach. *Canadian Journal of Civil Engineering* Vol. 39, no. 4, 2012, pp. 448-459.

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, Vol.1855, pp.160–167, 2003.

Chen C., K.Petty, A. Skabardonis, P. Varaiya, and Z. Jia. Freeway Performance 27 Measurement System: mining loop detector data. Transportation Research Record: Journal of the Transportation Research Board, Vol.1748, pp. 96–102, 2001.

Chen, L., and A. May. Traffic Detector Errors and Diagnostics. Transportation Research Record: Journal of the Transportation Research Board, No. 1132. Transportation Research Board of the National Academies, Washington, D.C., 1987, pp. 82–93.

Chen, W., Guo, F., and Wang, F. Y. A survey of traffic data visualization. Intelligent Transportation Systems, IEEE Transactions on,16(6), 2970-2984, 2015.

Cheng, W. and Washington, S.. New Criteria for Evaluating Methods of Identifying Hot Spots. Transportation Research Record: Journal of the Transportation Research Board, 2008, 76-85.

Chien, S. I. J., and S. M. Chowdhury. Simulation-Based Estimates of Delays at Freeway Work Zones. Transportation Research Board 80th Annual Meeting, Washington D.C., CD-ROM, 2000.

Chiu, Y. C., and P. B. Mirchandani. Online Behavior–Robust Feedback Information Routing Strategy for Mass Evacuation. IEEE Transactions on Intelligent Transportation Systems, Vol. 9, No. 2, 2008, pp. 264–274.

Cohen, H., and F. Southworth. On the Measurement and Valuation of Travel Time Variability Due to Incidents on Freeways. Journal of Transportation and Statistics, Vol. 2 (2). 123-131, 1999.

Corey, J., Lao, Y., Wu, J. and Wang, Y., Detection and Correction of Inductive Loop Detector Sensitivity Errors Using Gaussian Mixture Models, Transportation Research Record: Journal of the Transportation Research Board, Vol. 2256, pp.120-129, 2011.

Dahlgren, J., In What Situations Do High Occupancy Vehicle Lanes Perform Better Than General Purpose Lanes. University of California Transportation Center. 2001.

Deacon, J. A., Zegeer, C. V., and Deen, R. C., Identification of Hazardous Rural Highway Locations. Transportation Research Record, 543, pp.16-33, 1975.

Dijkstra, E. W.. A note on two problems in connexion with graphs. *Numerische mathematic 1*, no. 1, 1959, pp. 269-271.

Dutzik, T., and R. Pregulman (2003). "More Roads, More Traffic: Why Highway Constructions Won't Solve Traffic Congestion in Washington." WashPIRG Foundation.

Du, F., Brule, J., Enns, P., Manjunatha, V, and Segev, Y. MetroViz: Visual Analysis of Public Transportation Data, arXiv:1507.05215 [cs.HC], 2015

Federal Highway Administration, US Department of Transportation. 2014. Moving Ahead for Progress in the 21st Century [Online]. Available: https://www.fhwa.dot.gov/map21/ [Accessed August 26, 2014.

Ferraiolo, D.F. and Kuhn, D.R. 1992. Role-Based Access Control. 15th National Computer Security Conference: 554–563.

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.

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

Google Map APIs, accessed, https://developers.google.com/maps/ on April 5, 2016.

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

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

Hallenbeck, M.E., J.M. Ishimaru, and J. Nee (2003). Measurement of Recurrin versus Non-Recurring Congestion: Technical Report. Research report for Washington State Department of Transportation. WA-RD 568.1. Seattle.

Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning; Data Mining, Inference, and Prediction, second ed. New York, NY: Springer Verlag,

Hart, P. E., Nilsson, N. J., and Raphael, B.. A formal basis for the heuristic determination of minimum cost paths. Systems Science and Cybernetics, *IEEE Transactions on 4*, no. 2, 1968, pp. 100-107.

Hauer, E. 1996. Identification of Sites with Promise. Transportation Research Record, 1542, 54-60.

Hauer, E., Harwood, D. W., Council, F. M. & Griffith, M. S. 2002. Estimating Safety by the Empirical Bayes Method - A Tutorial. Transportation Research Record, 1784, 126-131.

Haworth, J., and Cheng T. Non-parametric regression for space–time forecasting under missing data. *Computers, Environment and Urban Systems*, Vol. 36, no. 6, 2012, pp. 538-550.

Henrickson, Kristian Carl. "Flexible and Robust Treatments for Missing Traffic Sensor Data." Masters Thesis., 2014.

Henrickson, Kristian, Yajie Zou, and Yinhai Wang. "Flexible and Robust Method for Missing Loop Detector Data Imputation." Transportation Research Record: Journal of the Transportation Research Board 2527 (2015): 29-36.

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

Hoogendoorn, S.P., and P. Bovy. "Pedestrian route-choice and activity scheduling theory and models." Transportation Research Part B: Methodological 38, no. 2, 2004, pp. 169-190.

Horton, N. J., and Lipsitz, S. R. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, Vol. 55, no. 3, 2001, pp. 244-254.

Hosp, A. K. U.S. Patent No. 9,092,828. Washington, DC: U.S. Patent and Trademark Office, 2015.

Huang, X., Zhao, Y., Yang, J., Zhang, C., Ma, C., and Ye, X.. TrajGraph: A Graph-Based Visual Analytics Approach to Studying Urban Network Centralities Using Taxi Trajectory Data. Visualization and Computer Graphics, IEEE Transactions on, 22(1), 160-169, 2016.

INRIX, INRIX Traffic Data Services: Tapping into Real-time Traffic Flow, Accessed online http://cta.ornl.gov/TRBenergy/trb_documents/2012_presentations/192_%20T%20Trepanier%20-%20INRIX%20LeeSchipper%20Data%20Workshop%20TRB2012.pdf on August, 14, 2013.

Khazaei, H., Zareian, S., Veleda, R., and Litoiu, M. Sipresk: A Big Data Analytic Platform for Smart Transportation. In EAI International Conference on Big Data and Analytics for Smart Cities, 2015.

Klein, L., M K. Mills and D. R.P. Gibson. Traffic Detector Handbook: Third Edition. Publication FHWA-HRT-06-108. FHWA, U.S. Department of Transportation, 2006.

Kloeckl, K., Chen, X., Sommer C., Ratti C., and Biderman A. Trains of data. [Online]. Available: http://senseable.mit.edu/trainsofdata/, on April 5, 2016.

Kong, D. D. Gray, and H. Tao (2006). A Viewpoint Invariant Approach for Crowd Counting. Proceedings of the 18th International Conference on Pattern Recognition, Vol. 3, 1187-1190.

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.

Lao, Y., Zhang, G., Wang, Y., & Milton, J. 2014. Generalized nonlinear models for rear-end crash risk analysis. Accident Analysis & Prevention, 62, 9-16.

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

Li, J., C. J. Lan, and X. Gu (2006). Estimation on Incident Delay and Its Uncertainty on Freeway Networks. Transportation Research Board 85th Annual Meeting, Washington D.C., CD-ROM.

Li, L., Li, Y., and Li, Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies* Vol. 34, 2013, pp. 108-120.

Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2, no. 3 (2002): 18-22.

Lin, W. H., and F. Daganzo (1997). "A Simple Detection Scheme for Delay-Inducing Freeway Incidents." Transportation Research, Part A. Vol. 31(2). 141-155.

Little, R. J. A. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics* Vol. 6, no. 3, 1988, pp. 287-296.

Ma, S., Zheng, Y., & Wolfson, O. T-share: A large-scale dynamic taxi ridesharing service. In Data Engineering (ICDE), 2013 IEEE 29th International Conference on (pp. 410-421), 2013.

Ma, X., McCormack, E., and Wang, Y., "Processing Commercial GPS 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.

Ma, X., Wu Y., and Wang. Y., "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.

Ma, X., Yu, R., and Wang, Y., "Developing a Regional Map-Based Platform for Spatial and Temporal Assessment of Traffic Emission Inventory", Proceedings of COTA International Conference of Transportation Professionals (CICTP). Aug. 2012.

Malinovskiy, Y., Y. Wu and Y. Wang. Video-Based Monitoring of Pedestrian Movements at Signalized Intersections, In Transportation Research Record: Journal of the Transportation Research Board. Washington, DC., Vol. 2073, pp. 11-17, 2008.

Meinshausen, Nicolai. Quantile regression forests. The Journal of Machine Learning Research 7 (2006): 983-999

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.

Mountain, L., Fawaz, B. & Jarrett, D. 1996. Accident Prediction Models for Roads with Minor Junctions. Accident Analysis & Prevention, 28, 695-707.

Morioka, M., Kuramochi, K., Mishina, Y., Akiyama, T. , and Taniguchi, N. City Management Platform Using Big Data from People and Traffic Flows. Hitachi Review, 64(1), 53, 2015.

Ni, D., Leonard, J. D., Guin, A., and Feng, C. Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. *Journal of transportation engineering* Vol. 131, no. 12, 2005a, pp. 931-938.

Ni, D., and Leonard, J. D. Markov chain monte carlo multiple imputation using bayesian networks for incomplete intelligent transportation systems data. *In Transportation Research Record: Journal of the Transportation Research Board* No. 1935-1, Transportation Research Board of the National Academies, Washington, D.C., 2005b, pp. 57-67.

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

Olmstead, T. (1999). Pitfall to Avoid When Estimating Incident-Induced Delay by Using Deterministic Queuing Models. Transportation Research Record 1683, pp. 38-46. Transportation Research Board. Washington D.C.

OneBusAway APIs, accessed, http://developer.onebusaway.org/modules/onebusaway-application-modules/1.1.14/api/where/index.html, on April 5, 2016.

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

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

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

Pack, M. L., Bryan, J. R., and Steffes, A. Overview and Status of Regional Integrated Transportation Information System in the National Capital Region. In Transportation Research Board 87th Annual Meeting (No. 08-1299), 2008.

Persaud, B., Lyon, C. & Nguyen, T. 1999. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. Transportation Research Record: Journal of the Transportation Research Board, 1665, 7-12.

Petty, K., Kwon, J., & Skabardonis, A. Arterial Performance Measurement System—APeMS. In Workshop on Performance Measures, 85th TRB Annual Meeting. CD-ROM. Washington DC, 2005.

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

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

Puget Sound Regional Council (PSRC) (2005). Growth in Traffic and Vehicle Miles Traveled. http://www.psrc.org/datapubs/pubs/trends/t2aug05.pdf. Accessed on July 5, 2006.

Pu, J., Liu S., Ding Y., Qu H., and Ni L. T-Watcher: A new visual analytic system for effective traffic surveillance, in Proc. IEEE 14th International Conference on MDM, vol. 1, pp. 127–136, 2013.

Qu, L., Li L., Zhang, Y., and Hu, J. PPCA-based missing data imputation for traffic flow volume: a systematical approach. *Intelligent Transportation Systems, IEEE Transactions on* Vol. 10, no. 3 2009, pp. 512-522

Rajagopal, R., and P. P. Varaiya. Health of California's Loop Detector System (UCB-ITS-PRR-2007-13). California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2007.

Ritchie, S. G. and R. L. Cheu. Simulations of Freeway Incident Detection Using Artificial Neural Networks. Transportation Research C, Vol. 1(3): 203-217, 1993.

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

Rubin, D. B. Inference and missing data. *Biometrika*, Vol. 63, no. 3, 1976, pp. 581-592.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2014.

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

Sandhu, R., Coyne, E.J., Feinstein, H.L. and Youman, C.E. 1996. Role-Based Access Control Models. IEEE Computer (IEEE Press) 29 (2): 38–47.

Schafer, J. L., and Graham, J. W. Missing data: our view of the state of the art. *Psychological methods* Vol. 7, no. 2, 2002, pp. 147.

Schafer, J. L. Analysis of incomplete multivariate data. CRC press, 2010.

Scheepens, R., Hurter, C., Van De Wetering, H., and Van Wijk, J. J. Visualization, selection, and analysis of traffic flows. Visualization and Computer Graphics, IEEE Transactions on, 22(1), 379-388, 2016.

Shekhar, Shashi, C. Tien Lu, R. Liu, and C. Zhou. CubeView: a system for traffic data visualization. In Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on, pp. 674-678. IEEE, 2002.

Siemens, Precise measurement and intelligent processing of travel time data, accessed https://www.mobility.siemens.com/mobility/global/SiteCollectionDocuments/en/road-solutions/urban/infrastructure/travel-time-measure-en.pdf on Mar. 3, 2016

Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." Bioinformatics 28, no. 1 (2012): 112-118.

Sun, J., Peng, Z. R., Shan, X., Chen, W., and Zeng, X., Development of web-based transit trip-planning system based on service-oriented architecture, *Transportation Research Record:*

Journal of the Transportation Research Board, No. 2217, Transportation Research Board of the National Academies, Washington, DC, pp.87-94, 2011.

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y. J., and Li F. 2013. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies*, Vol. 28, pp. 15-27.

Teng, H and Y. Qi. (2003). Application of Wavelet Technique for Freeway Incident Detection. Transportation Research C, Vol. 11, No. 3: 289-308.

Transportation Research Board. (2000). Highway Capacity Manual. National Research Council. Washington D.C.

Tsekeris, Theodore, and Alexander Skabardonis. On-line performance measurement models for urban arterial networks. Transportation Research Board Annual Meeting, Vol. 935, 2004.

Tu, H., Van Lint, J., & Van Zuylen, H. (2007). Impact of traffic flow on travel time variability of freeway corridors. Transportation Research Record: Journal of the Transportation Research Board, (1993), 59-66.

Uber APIs, accessed, https://developer.uber.com/, on April 5, 2016.

Van Buuren, S., and Oudshoorn, K. Flexible multivariate imputation by MICE. Leiden, The Netherlands: TNO Prevention Center (1999).

Vlahogianni, E. I., Park, B. B., and van Lint, J. W. C. Big data in transportation and traffic engineering. Transportation Research Part C: Emerging Technologies, 58, 161, 2015.

Wang, Z., Ye, T., Lu, M., Yuan, X., Qu, H., Yuan, J., & Wu, Q. Visual exploration of sparse traffic trajectory data. Visualization and Computer Graphics, IEEE Transactions on, 20(12), 1813-1822, 2014.

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.

Washington State Department of Transportation, 2014. WSDOT's Handbook for corridor Capacity Evaluation.

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

WSDOT, Corridor Capacity Report, accessed http://www.wsdot.wa.gov/Accountability/Congestion/2015.htm on Mar. 3, 3016.

WSDOT, Navigate the Gray Notebook, accessed http://www.wsdot.wa.gov/Accountability/GrayNotebook/navigateGNB.htm on August, 19, 2013.

WSDOT, The 2012 Congestion Report, accessed http://wsdot.wa.gov/publications/fulltext/graynotebook/CR12.pdf on August, 19, 2013.

WSDOT, Travel Time Details, accessed http://www.wsdot.wa.gov/Traffic/seattle/questions/traveltimesdetail.htm on Mar. 3, 2016.

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

Xie, Gang, and Brian Hoeft. Freeway and Arterial System of Transportation Dashboard: Web-Based Freeway and Arterial Performance Measurement System. Transportation Research Record: Journal of the Transportation Research Board 2271, 45-56, 2012.

Yuan, J., Zheng, Y., Zhang, L., Xie, X., and Sun, G. Where to find my next passenger. In Proceedings of the 13th international conference on Ubiquitous computing (pp. 109-118). ACM, 2011.

Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., and Chen, C. Data-driven intelligent transportation systems: A survey. Intelligent Transportation Systems, IEEE Transactions on, 12(4), 1624-1639, 2011.

Zhang, X., Y. Wang, and N. L. Nihan. Investigating Dual-Loop Errors Using Video Ground-Truth Data. Proc., ITS America 2003 Annual Meeting, Minneapolis, Minn., 2003.

Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., and Chang, E. Diagnosing New York city's noises with ubiquitous data. InProceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 715-725). ACM, 2014.

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.

Zou, Y., Wu, L. & Lord, D. 2015. Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in Negative Binomial models. Analytic Methods in Accident Research, 5–6, 1-16.