DEVELOP A REGIONAL MULTI-SOURCE DATABASE SYSTEM FOR SAFETY DATA MANAGEMENT AND ANALYSIS IN RITI COMMUNITIES IN WASHINGTON STATE

FINAL PROJECT REPORT

by

Yinhai Wang, Ziqiang Zeng, Christopher Gottsacker, and Hao (Frank) Yang

University of Washington

for
Center for Safety Equity in Transportation (CSET)

USDOT Tier 1 University Transportation Center

University of Alaska Fairbanks

ELIF Suite 240, 1764 Tanana Drive

Fairbanks, AK 99775-5910

April 19, 2019

In cooperation with U.S. Department of Transportation,
Research and Innovative Technology Administration (RITA)

INE/CSET 19.11

CSET

rural • isolated • tribal • indigenous

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Center for Safety Equity in Transportation, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle Develop a Regional Multi-Source Database System for Safety Data Management and Analysis in RITI Communities in Washington State	**5. Report Date** April 19, 2019 **6. Performing Organization Code**
7. Author(s) and Affiliations Yinhai Wang, Ziqiang Zeng, Christopher Gottsacker, and Hao (Frank) Yang University of Washington	**8. Performing Organization Report No.** INE/CSET 19.11
9. Performing Organization Name and Address Center for Safety Equity in Transportation ELIF Building Room 240, 1760 Tanana Drive Fairbanks, AK 99775-5910	**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.**
12. Sponsoring Organization Name and Address United States Department of Transportation Research and Innovative Technology Administration 1200 New Jersey Avenue, SE Washington, DC 20590	**13. Type of Report and Period Covered** Final report, Sep 2018 – Feb 2019 **14. Sponsoring Agency Code**

15. Supplementary Notes

Report uploaded to:

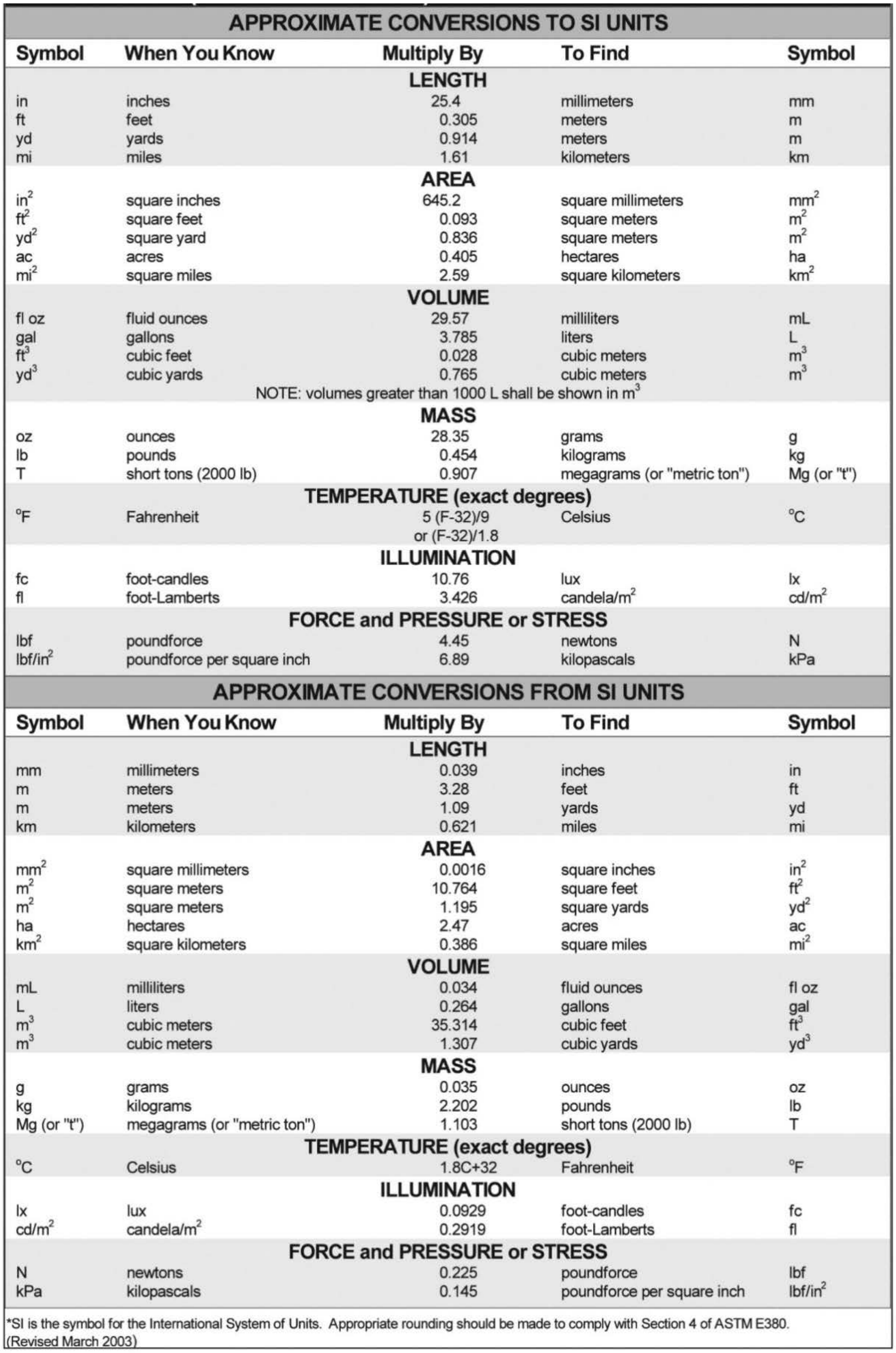
16. Abstract

Rural, Isolated, Tribal, and Indigenous (RITI) communities across the United States are disadvantaged from a transportation safety perspective. Particular concern is focusing on rural road safety. Since RITI communities often do not have the capability and resources to sufficiently solve roadway safety problems, several challenges are encountered for addressing transportation safety issues in RITI communities, including: (1) Crashes are often distributed along roads in RITI areas without known patterns; (2) Strategies to address safety issues are diverse for different RITI communities and draw from several safety areas. As a result, there is a critical need to realize equitably-augmented safety solutions that address the needs of these underserved and underinvested RITI communities. To address this gap, this project aims to develop a regional multi-source database system for traffic safety data management and analysis of RITI communities in Washington State. The existing crash data sources in RITI communities in Washington was identified and documented. The crash data on rural routes was extracted from the raw data from Washington State Department of Transportation and integrated into the multi-source database system, including traffic flow characteristics, crash attributes and contribution factors, crash-related trauma data and medical records, weather conditions, etc. The Colville tribe also provided the crash data in their tribal communities under a confidentiality agreement. A multi-source database fusion and integration system architecture was designed. Microsoft SQL Server 2012 was used to implement the database and manage the data. A six-step data quality control method was employed to clean the data by wiping out the outliers from spatial and temporal aspects. The tribal crash data was made accessible to authorized users so they can download the datasets by using password, while the WSDOT crash data was set to be public for all the users. A safety analysis module was developed for visualizing the data in the regional multi-source database system in RITI communities. The data visualization platform is developed based on the Vaadin Framework. The users can interact with the interface for data analysis. A safety performance index and a potential safety improvement index were also developed. By combining the two indexes, one can easily identify crash hotspots and the key influencing factors to consider in an improvement package.

17. Key Words Baseline data; Multi-source database system; Transportation; Safety data management and analysis; ; RITI communities	**18. Distribution Statement**

19. Security Classification (of this report)	**20. Security Classification (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified.	Unclassified.	32	N/A

SI* (MODERN METRIC) CONVERSION FACTORS



APPROXIMATE CONVERSIONS TO SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^{o}F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^{o}C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^{o}C	Celsius	1.8C+32	Fahrenheit	^{o}F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Rural, Isolated, Tribal, and Indigenous (RITI) communities across the United States are disadvantaged from a transportation safety perspective. The RITI communities in Washington State include twenty-nine federally recognized Tribes. Fatality Analysis Reporting System (FARS) data from 2002 through 2011 shows the traffic fatality rate for Native Americans is 3.9 times higher than for non-Native Americans in Washington. Two-thirds of Native American pedestrian fatalities within Washington occurred in rural areas. Inadequate or non-existent bus systems increase the number of pedestrians on tribal lands. Some Tribes have non-contiguous lands with housing and services on separate assets. Many communities have few or no sidewalks, marked crosswalks or street lights. Chronic underfunding of traffic safety initiatives and related programs plays a significant role in these disproportionately disadvantaged RITI areas. There is a strong need to move toward equitably augmented safety solutions to satisfy the needs of these underrepresented groups in Washington State. As a result, it is critical to develop a data-driven safety related analytical platform to investigate the risk factors and safety status associated with RITI communities. However, there is no recognized comprehensive database for RITI communities in Washington State. Without high-quality data and data management systems, RITI communities may not able to justify their needs to receive funding for safety improvements. It is necessary to build up the comprehensive data infrastructure to enhance the ability to develop informed data-driven plans and mitigation strategies. To address this gap, this project aims to develop a regional multi-source database system for traffic safety data management and analysis of RITI communities in Washington State. This research effort will gather and leverage existing traffic crash databases within Washington State and develop a database system to dynamically retrieve rural traffic crash data and graphically visualize the data for crash attribute analysis. As part of baseline crash data infrastructure establishment, the proposed data platform can enable effective traffic safety program management at all levels in RITI communities, and design and implement appropriate countermeasures to mitigate rural crash severities and risks.

Surveying the existing related web-based archived data user service systems was an important first step. The research team found that while various web-based archived data user service systems have been developed during the past decade, most of them focus on travel time analysis, and little existing effort concentrated on multi-source safety data fusion and integration for RITI communities. The existing crash data sources in RITI communities in Washington were identified and documented. The crash data on rural routes were extracted from the raw data provided by Washington State Department of Transportation (WSDOT) and integrated into the multi-source database system, including traffic flow characteristics, crash attributes and causal factors, crash-related trauma data and medical records, weather conditions, etc. The Colville tribe also provided the crash data in their tribal communities under a confidentiality agreement. A multi-source database fusion and integration system architecture was designed. Microsoft SQL Server 2012 was used to implement the database and manage the data. A six-step data quality control method was employed to clean the data by wiping out the outliers from spatial and temporal aspects. The tribal crash data were made accessible to only authorized users so they can download the datasets by using a password. A safety analysis module was developed for visualizing the data in the regional multi-source database system in RITI communities. The data visualization platform is developed based on the Vaadin Framework. The users can interact with the interface for data analysis. A safety performance index (SPI) and a potential safety improvement index (PSII) were also developed. By combining the two indexes on the regional map, one can easily identify crash hotspots and the key influencing factors to consider in safety improvement packages.

CHAPTER 1. INTRODUCTION

1.1. Research Background

Traffic crashes cost billions of dollars in life loss and property damage annually across the United States. In 2000, Washington State adopted a strategic highway safety plan Target Zero — a goal to eliminate traffic fatalities and serious injuries on Washington's roadways to zero by 2030 (Washington Traffic Safety Commission, 2016). However, great challenges must be addressed in minimizing fatalities and serious injuries in Rural, Isolated, Tribal, and Indigenous (RITI) communities. RITI communities across the United States are disadvantaged from a transportation safety perspective. One particular concern is rural road safety. Official data from Federal Highway Administration (FHWA) show that, in 2012, 54 percent of all fatalities occurred on rural roads while only 19 percent of the US population lived in rural communities (Federal Highway Administration, 2012). As shown in Figure 1, the fatality rate was 2.4 times higher in rural areas than in urban areas (1.81 and 0.74, respectively) (Federal Highway Administration, 2014). Since RITI communities often do not have the capability and resources to sufficiently solve roadway safety problems, several challenges are encountered for addressing transportation safety issues in RITI communities, including: (1) Crashes are often randomly distributed on local and rural roads in RITI areas; (2) Strategies to address safety issues are diverse for different RITI communities and draw from several safety areas (Federal Highway Administration, 2016).

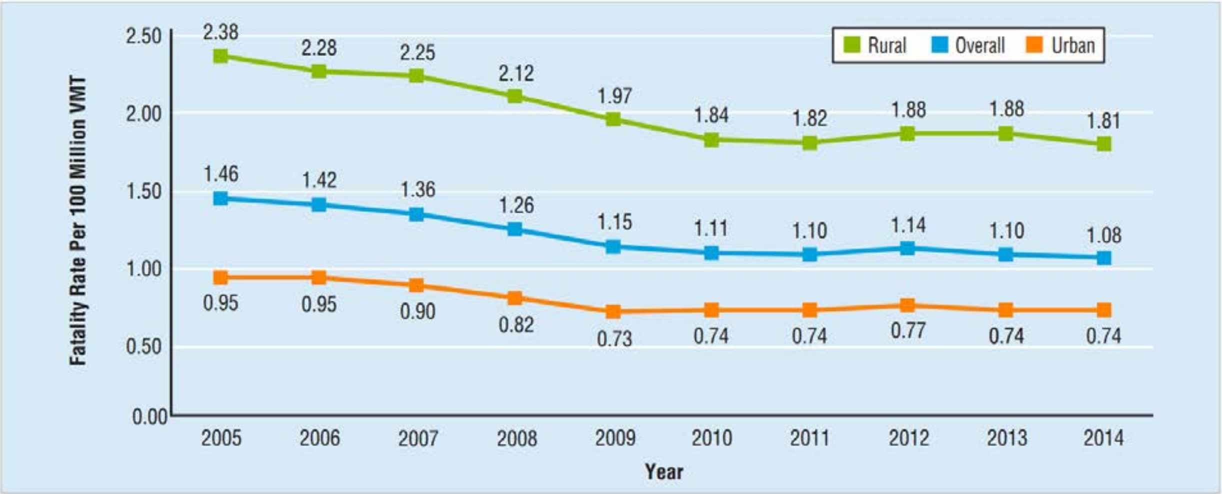


Figure 1 Fatality Rates per 100 Million Vehicle Miles Traveled, by Year and Location, 2005‑2014 (Federal Highway Administration, 2014). Source: FARS 2005-2013 Final File, 2014 ARF; VMT ‑ Federal Highway Administration

To meet the transportation safety needs of RITI communities, Washington State also faces many challenges. Twenty-two percent of the state's major rural locally and state-maintained roads are in poor condition. An additional 52 percent of rural roads are in mediocre or fair condition. Between 2002 and 2011, 61% of traffic fatalities occurred on rural roads, even though many more miles are traveled on urban roads in Washington State (Washington Traffic Safety Commission, 2013). Figure 2 illustrates the rural and urban fatality rates based on data from 2002 through 2011 in Washington. The fatality rate on Washington's rural non-Interstate roads was 1.76 fatalities per 100 million vehicle miles of travel in 2013, nearly three and a half times higher than the 0.52 fatality rate on all other roads and highways in the state (TRIP, 2015). According to the data from Washington State Strategic Highway Safety Plan 2016, more than half (52%) of impairment-involved fatalities occurred in rural areas during 2012-2014, and unrestrained occupants are also more likely to die in rural road crashes (Washington Traffic Safety Commission, 2016). It is obvious that rural roadway safety has become an important social issue influencing the sustainable development of the RITI communities in Washington State. The greatest challenge in addressing fatalities and serious injuries on rural highways is the geographic randomness of collisions scattered over tens of thousands of miles. Unlike on urban roads, there are few concentrations of serious crashes, and the locations of crashes are not consistent from year to year. As a result, identifying the best approaches for behavioral improvements and safety infrastructure enhancement can be difficult.

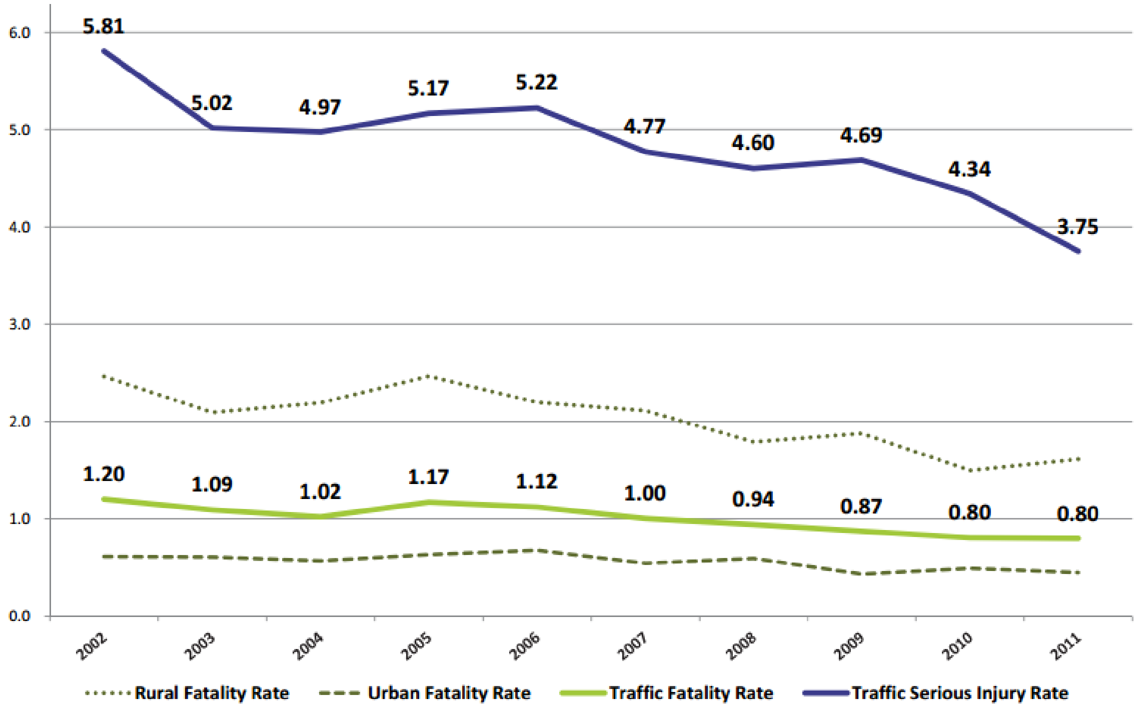
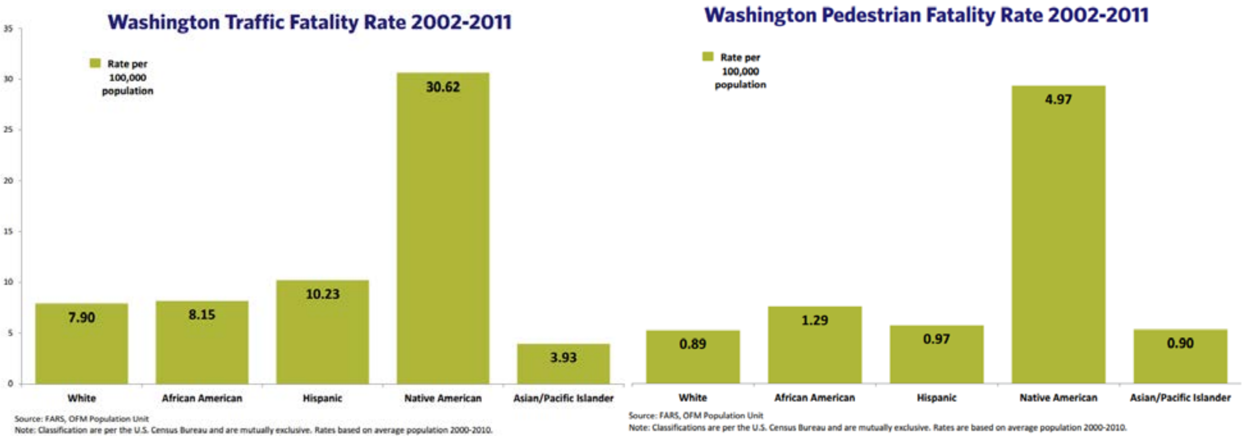


Figure 2 Traffic fatality and serious injury rates (2002-2011) per 100 million vehicle miles traveled in Washington state (Washington Traffic Safety Commission, 2016).

The RITI communities in Washington State include twenty-nine federally recognized Tribes. Fatality Analysis Reporting System (FARS) data from 2002 through 2011 shows the traffic fatality rate for Native Americans is 3.9 times higher than for non-Native Americans in Washington. Two-thirds of Native American pedestrian fatalities within Washington boundaries occurred in rural areas (Washington Traffic Safety Commission, 2013). Figure 3 shows the comparison results of Washington pedestrian and traffic fatality rates between Native Americans and non-Native Americans from 2002 through 2011. Inadequate or non-existent bus systems increase the number of pedestrians on tribal lands. Some Tribes have non-contiguous lands with housing and services on separate assets. Many communities have few or no sidewalks, marked crosswalks or street lights. Chronic underfunding of traffic safety initiatives and related programs plays a significant role in these disproportionately disadvantaged RITI areas. There is a critical need to move toward equitably-augmented safety solutions to satisfy the needs of these underrepresented groups in Washington State.



Source: FARS, OFM Population Unit
Note: Classification are per the U.S. Census Bureau and are mutually exclusive. Rates based on average population 2000-2010.

Source: FARS, OFM Population Unit
Note: Classifications are per the U.S. Census Bureau and are mutually exclusive. Rates are based on average population 2000-2010.

Figure 3 Comparison results on pedestrian and traffic fatality rates between Native Americans and non-Native Americans in Washington from 2002 through 2011 (Washington Traffic Safety Commission, 2013).

1.2. Problem Statement

Seven high-risk factors that result in fatalities and serious injuries occurring on reservation roads include driver impairment, lane departure, unrestrained vehicle occupants, intersections, young drivers (aged 16-25), speeding, and unlicensed drivers based on the data from 2012 through 2014 (Washington Traffic Safety Commission, 2016). As a result, it is critical to develop a data-driven safety related analytical platform to investigate the risk factors and safety status associated with RITI communities. However, existing databases in Washington State are incomplete for RITI communities. A survey made by National Association of Counties (NAC) in 2009 revealed that only 42 percent of counties surveyed maintained a database that tracks the number and types of crashes on their rural roads and slightly under half of the respondents have conducted a road safety audit (National Association of Counties, 2009). Existing databases are still incomplete for most of the RITI communities. Without high-quality data and data management systems, RITI communities are not eligible to receive funding for safety improvements. It is necessary to build up the comprehensive data infrastructure to enhance the ability to develop informed data-driven plans and mitigation strategies. To address this gap, this project aims to develop a regional multi-source database system for traffic safety data management and analysis of RITI communities in Washington State. This research effort will gather and leverage existing traffic accident databases within the Washington State and develop a database system to dynamic retrieve rural traffic crash data and graphically visualize the data for crash attribute analysis. As part of baseline crash data infrastructure establishment, the proposed data platform can enable effective traffic safety program management at all levels in RITI communities, and design and implement appropriate countermeasures to mitigate rural crash severity and risk.

1.3. Research Objectives and Tasks

The research objectives are as follows:

- Identify and document existing crash data sources in RITI communities in Washington;
- Fuse and integrate region-wide RITI community safety-related baseline datasets in Washington, including traffic flow characteristics, crash attributes and contributing factors, crash-related trauma data and medical records, weather conditions, etc.;
- Design and implement the data platform and its supporting relational database, such as SQL database to unify data storage and management;
- Develop methods for RITI community safety data quality control; and
- Develop safety analysis and high risk location identification methods with user-friendly interfaces on the data platform.

Based on the above research objectives, five tasks were identified and put in the work plan of this project as below:

- Task 1: Review the existing multi-source safety database integration systems, and collect multiple-year rural crash data, traffic sensor data, and weather conditions data from WSDOT and related RITI communities in Washington;
- Task 2: Design the database and fuse multi-source datasets for safety performance evaluation in RITI communities;
- Task 3: Develop the regional database system to analyze, visualize, and manage multi-source safety data;
- Task 4: Examine and validate the database system and perform comprehensive safety performance analysis;
- Task 5: Draft and finalize project report.

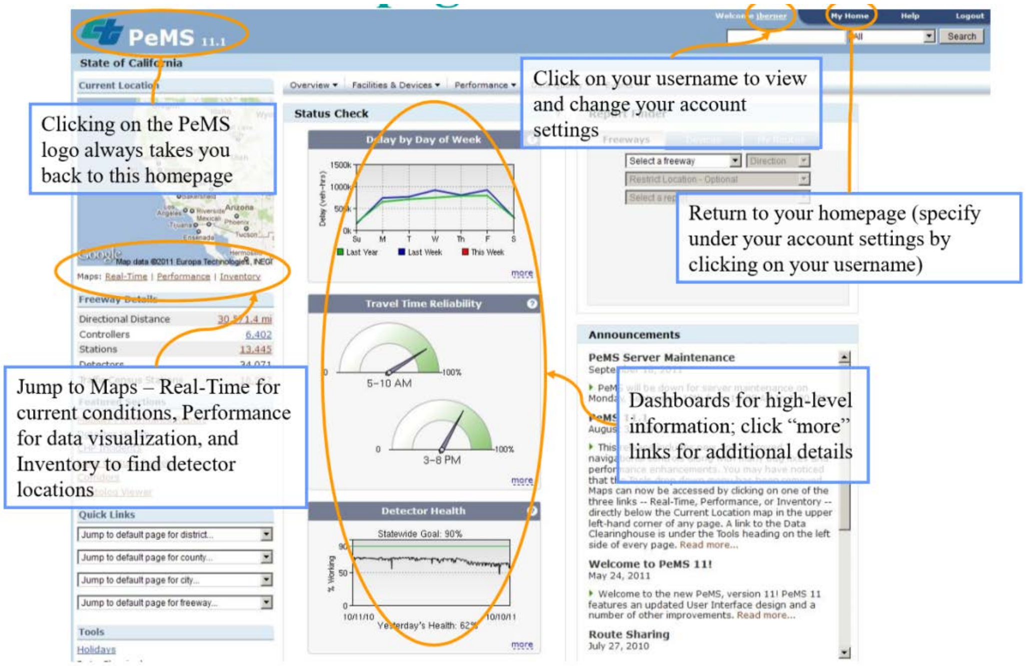
These research tasks can address CSET baseline data needs through the following aspects: 1) Gather and integrate region-wide multiple-year RITI community safety-related baseline data; 2) Design and implement online data platform and its supporting relational database, such as SQL database to unify data storage and management; and 3) Develop methods for RITI community safety data quality control and cleaning. The developed database fuses and integrates multiple source data including traffic flow characteristics, crash attributes and contributing factors, crash-related trauma data and medical records, weather conditions, etc., to enable advanced safety enhancement studies.

CHAPTER 2. LITERATURE REVIEW

Various web-based archived data user service systems have been developed during the past decade. Most of them focus on travel time analysis and little existing effort has been concentrating on multi-source safety data fusion and integration for RITI communities. The related literature is classified into three categories to introduce the current state-of-the-art and practice: performance measurement system, multiple-agency system, and online data-driven analysis system.

2.1. Performance Measurement System

An early example is the online Freeway Performance Measurement System (PeMS) developed by the University of California, Berkeley since 1997, which is capable of analyzing freeway traffic sensor data and providing real-time performance measures, including travel times (Chen, 2003). PeMS was sponsored by the California Department of Transportation (Caltrans) and can develop tools and reports for traffic engineers, operators, and planners. It obtains data such as speed, vehicle-hours of delay, vehicle-miles traveled, and travel time statistics, from automatic sensors which are installed on most of the freeways in California. These data are integrated on a web interface and summarized in reports. It is easy for the policy makers to use PeMS to assess the effectiveness of their decisions and make adjustments; the traffic planners use it to monitor the congestion conditions and respond with countermeasures; the transportation engineers use the detailed data to improve conditions at specific locations; the travelers utilize the information for making more informed decisions; the researchers also use the database to analyze travel behavior. It is a data visualization and analysis tool which enables system monitoring and evaluation (See Figure 4). The visualization and analytical functions include travel time reliability, delay by day of week, detector health, and so on.



Clicking on the PeMS logo always takes you back to this homepage

Click on your username to view and change your account settings

Return to your homepage (specify under your account settings by clicking on your username)

Jump to Maps – Real-Time for current conditions, Performance for data visualization, and Inventory to find detector locations

Dashboards for high-level information; click "more" links for additional details

Figure 4 Visualization and analytical functions of PeMS (Source: http://media2.planning.org/APA2012/Presentations/S203_Caltrans%20Performance%20Measurement%20System.pdf)

Based on the PeMS, Petty et al. (2006) developed an arterial performance measurement system to estimate travel time on an arterial route using midblock (system) loop detectors. Ma (2008) developed a real-time performance measurement system for arterial traffic signals, which can automatically collect and archive high-resolution traffic signal data and generate a rich list of performance measures. Xie & Hoeft (2012) also built an integrated web-based freeway and arterial performance monitoring and measurement system, named Freeway and Arterial System of Transportation (FAST) dashboard. It has an intuitive web-based user interface to present real-time and historical freeway and arterial network monitoring and performance information which include most of those in the Highway Capacity Manual and other widely accepted professional handbooks. They also defined some new measures applied in this system such as delay volume.

Bertini et al. (2005) and Tufte et al. (2010) developed the Portland Oregon Regional Transportation Archive Listing (PORTAL) system for archiving and analyzing freeway data. Since 2004, it has received 20-s data from the 436 inductive loop detectors and other sensors in Portland area where the data archive includes transit data, freeway incident data, city traffic signal data, and truck weigh-in-motion data. The PORTAL 2.0 (Tufte et al., 2010) is the official transportation data archive for the Portland metropolitan area which provides a more-intuitive interface with improved data quality monitoring and control.

2.2. Multiple-Agency System

The researchers at the University of Maryland, College Park developed the Regional Integrated Transportation Information System (RITIS) (See Figure 5) which is a user-friendly multiple-agency system for sharing, disseminating, and archiving data (Pack et al., 2008). This database system integrates multiple data sources from different transportation agencies and focuses on freeway applications, including data fusion and standardization, and their relationship to data collection, regional transportation management, and so on. The participating agencies can view the regional traffic information and use it to enhance their operations. Figure 6 illustrates the interface of RITIS.

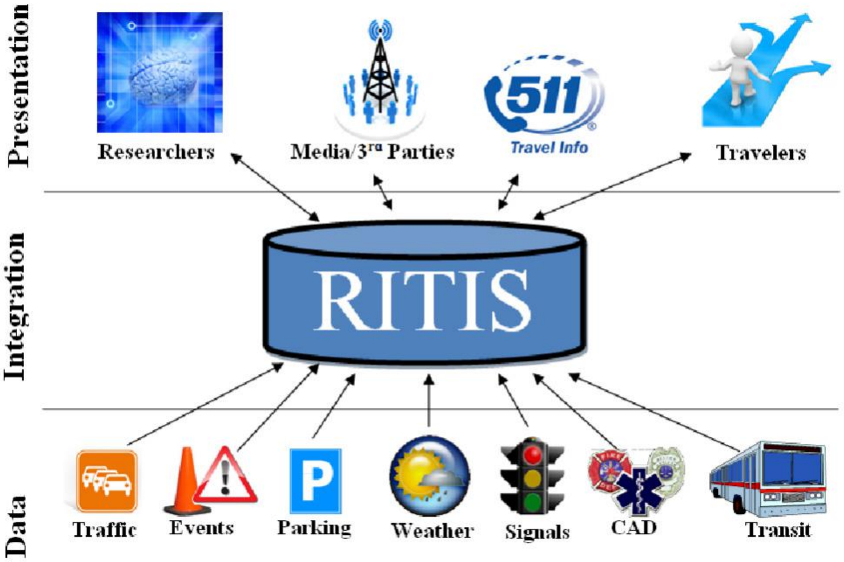


Figure 5 Framework of RITIS (Source: http://www.cattlab.umd.edu/?portfolio=ritis)

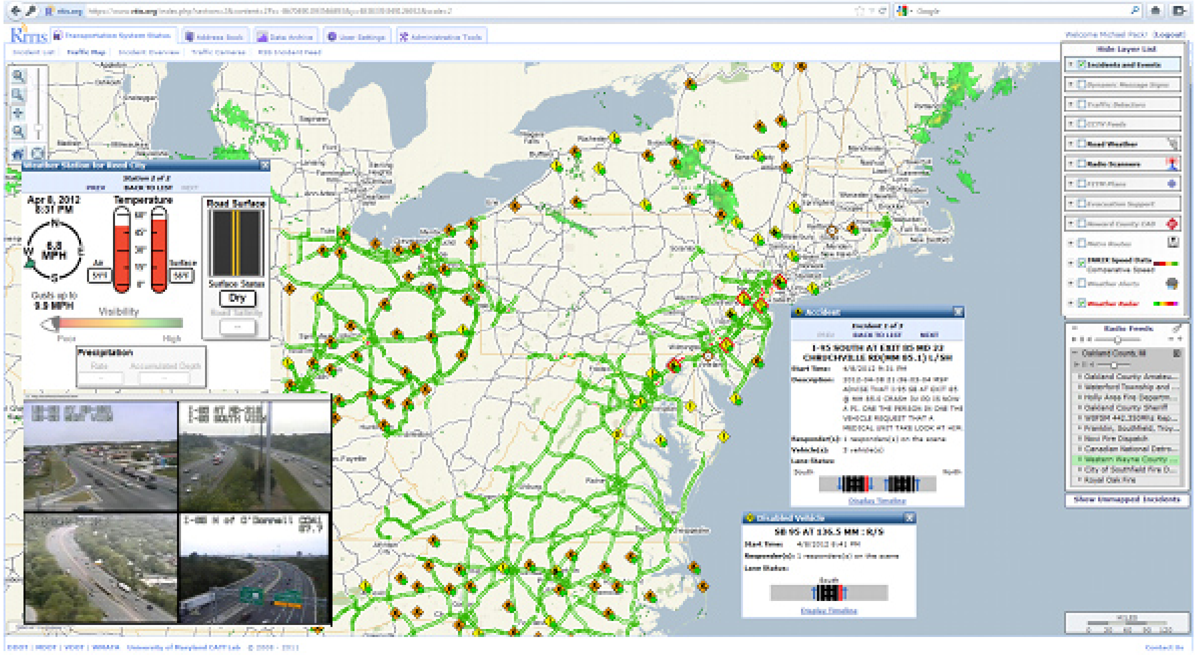
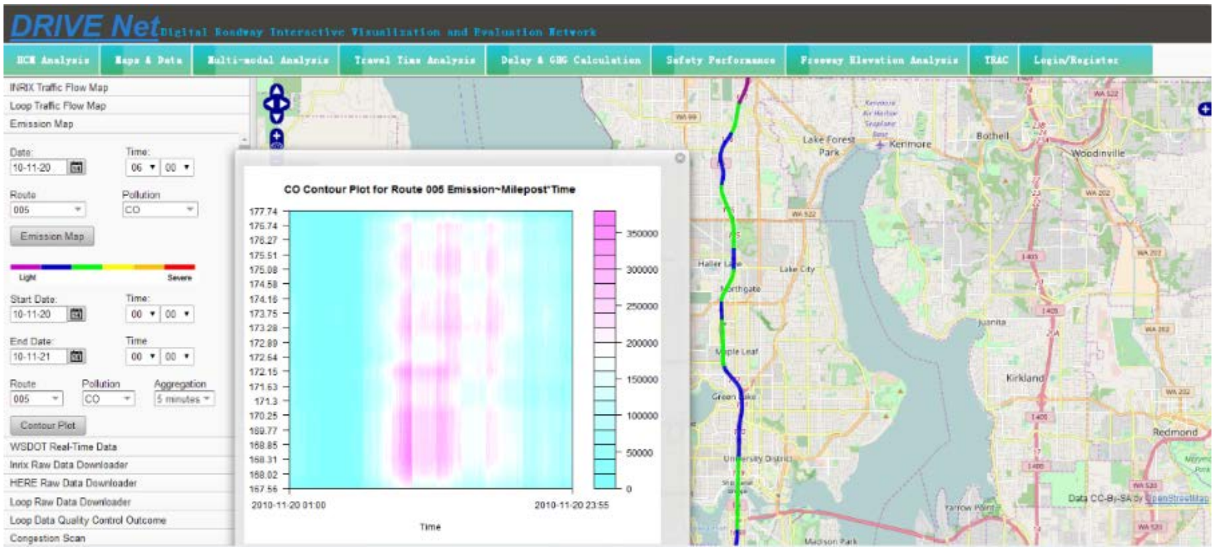


Figure 6 Interface of RITIS (Source: http://www.cattlab.umd.edu/?portfolio=ritis)

The University of Maryland has also developed various platforms for the visualization and analysis of transportation data, each designed for a specific purpose, such as incident analysis (Filippova & Pack, 2009). Wongsuphasawat et al. (2009) developed a novel, web-based, visual analytics tool named Fervor as an application which provides complicated but user-friendly analysis of transportation incident data sets. This tool has four featured visualization functions such as interactive maps, histograms, two-dimensional plots, and parallel coordinates plots. A rank-by-strength framework was also designed to quantify the strength of relationships among the different fields describing the data. Pack et al. (2005) developed a four-dimensional interactive visualization system for transportation management and traveler information which covers the major road networks of Washington D.C., Northern Virginia, and Maryland. The system can interact with real-time transportation databases to present animations of real-time traffic volume and speed along with incident data such as accident locations, lane closures, and responding agencies. It can also monitor and interact with the traffic control devices, sensors and even view the location of emergency response vehicles equipped with Global Positioning System (GPS) transceivers. VanDaniker & Pack (2009) developed a system named Transportation Incident Management Explorer (TIME) which can visualize real-time and historic traffic incident data. To be different from other systems, TIME can view the data related to an incident in a compact graphical overview. It can reduce the possibility of missing important information, enables the users to correlate different events, and speeds up the process of understanding the events which occur or occurred during the management of an incident. Lund & Pack (2010) also proposed web-based congestion and incident scanner tool which can provide dynamic and interactive analysis on traffic congestion conditions. The congestion performance is visualized for specific date ranges and locations in an intuitive and minimalistic interface. The users can interact with the visualizations to better identify the hotspots and easily correlate the congestion abnormalities with possible reasons.

2.3. Online Data-Driven Analysis System

Wang and his research team at the University of Washington (UW) Smart Transportation Applications and Research Laboratory (STAR Lab) developed a Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) system for visualizing real-time traffic conditions and performing online data-driven analysis of arterial traffic by using intersection loop and traffic signal timing data (Ma et al., 2011a; Wang et al., 2009; Ma et al., 2011b). DRIVE Net is a regional-wide web-based transportation decision system with multi-source data including traffic sensor, incident, accident, and travel time data, as shown in Figure 7. It can be regarded as a platform for transportation decision making optimization. While DRIVE Net has a safety analysis module (See Figure 8), it is only for urban area of Seattle rather than RITI communities.

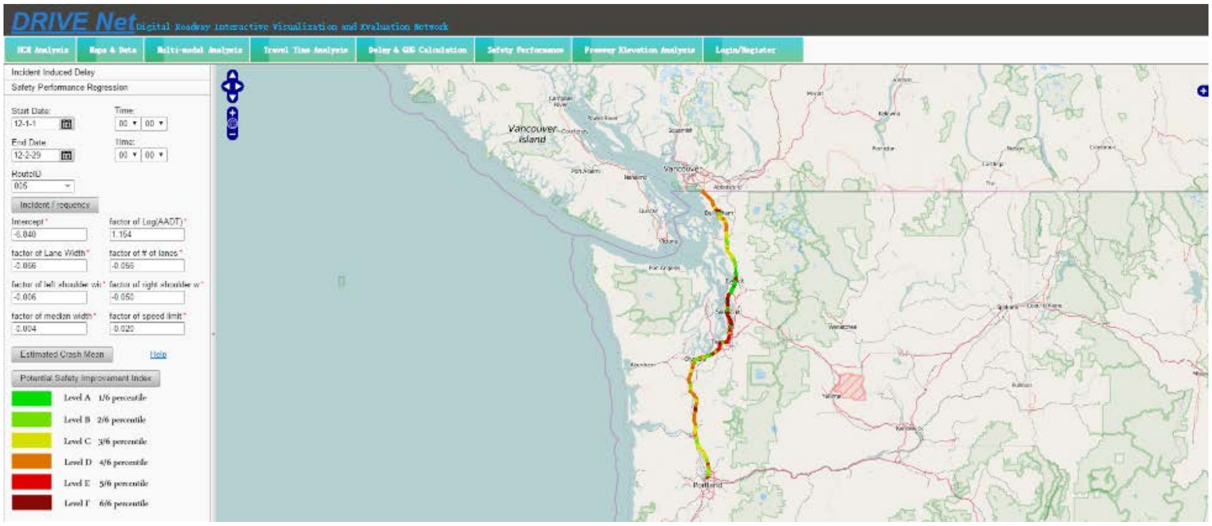


Date: 10-11-20
Time: 06 00
Route 005
Pollution CO

Start Date: 10-11-20
Time: 00 00
End Date: 10-11-21
Time 00 00
Route 005
Pollution CO
Aggregation 5 minutes

Figure 7 Interface of DRIVE Net (Source: http://uwdrive.net/STARLab)

Wu & Wang (2009) proposed a Google-map-based online system (Wu et al., 2007) for urban traveler information visualization and analysis. It has two functions, i.e., time-domain analysis and scatter plot analysis. There has been little existing effort concentrating on multi-source safety data fusion and integration for RITI communities. Wu et al. (2011) also developed a web-based analysis system for real-time decision support on arterial networks. The system is designed in a structure of four layers which include offline server, online server (middleware), online server (Java Servlet), and online client. This structure can distribute the computational burden on the server. The system can be used to monitor the arterial performance based on the loop detector data. These online data-driven analysis systems have not yet been applied to RITI communities.



Start Date: 12-1-1 Time: 00 ▾ 00 ▾

End Date: 12-2-29 Time: 00 ▾ 00 ▾

RouteID: 005

Intercept*: -6.840

factor of Log(AADT)*: 1.164

factor of Lane Width*: -0.066

factor of # of lanes*: -0.056

factor of left shoulder wi*: -0.006

factor of right shoulder w*: -0.050

factor of median width*: 0.004

factor of speed limit*: 0.020

🟩	Level A 1/6 percentile
🟩	Level B 2/6 percentile
🟨	Level C 3/6 percentile
🟧	Level D 4/6 percentile
🟥	Level E 5/6 percentile
🟥	Level F 6/6 percentile

Figure 8 Safety analysis module of DRIVE Net (Source: http://uwdrive.net/STARLab)

CHAPTER 3. MULTI-SOURCE DATABASE SYSTEM

3.1. Multi-Source Data Collection

In order to identify and document existing crash data sources in RITI communities in Washington, a two-fold strategy was implemented to collect the data from multiple sources: (1) Collect raw data from Washington State Department of Transportation (WSDOT) and extract the related crash data in the rural areas; (2) Collect safety data from tribal communities in Washington State. The details about the data collected from the different sources are summarized as follows.

3.1.1. Data from WSDOT

The WSDOT provided us crash data on the state routes from 2010 to 2016. The datasets include 266 attributes for each record such as collision report number, state route type (urban/rural), number of fatalities, number of injuries, number of pedal cyclists involved, number of pedestrians involved, number of motor vehicles involved, and so on. Table 1 summarizes the general information of the raw datasets. The total number of records for the seven years is 1,970,634, while the average number of records for each year is 281,519. The average numbers of fatalities, injuries, pedal cyclists involved, pedestrians involved, and motor vehicles involved are 1,367, 151,176, 3,467, 5,666, and 579,607, respectively, for each year. However, in order to identify the crash data for RITI communities, the data for the rural routes were extracted from the raw datasets. By using the attribute “state route type (urban/rural)”, it is not difficult to identify the datasets for rural routes. Table 2 shows the general information of the extracted datasets for rural routes.

Table 1 General information of the raw datasets.

Year	Total number of records	Total number of fatalities	Total number of injuries	Total number of pedal cyclists involved	Total number of pedestrians involved	Total number of motor vehicles involved
2010	267,612	1,230	150,698	3,320	5,634	546,281
2011	258,849	1,249	145,618	3,395	5,249	527,988
2012	262,419	1,159	145,713	3,299	5,543	534,891
2013	263,832	1,160	141,535	3,464	4,973	543,443
2014	285,118	1,318	150,548	3,561	5,835	590,478
2015	310,486	1,944	163,618	3,765	5,947	644,664
2016	322,318	1,507	160,502	3,467	6,481	669,506
Average	281,519	1,367	151,176	3,467	5,666	579,607
Total	1,970,634	9,567	1,058,232	24,271	39,662	4,057,251

The number of records on rural routes is 135,396 which is only 6.88% of the total records, and the average value for each year is 19,342. The average numbers of fatalities, injuries, pedal cyclists involved, pedestrians involved, and motor vehicles involved on the rural routes are 361, 13,176, 55, 155, and 33,712, respectively, for each year. Figure 9 illustrates the differences of the dynamic trends of the general information (i.e., number of records, number of fatalities, number of injuries, number of pedal cyclists involved, number of pedestrians involved, number of motor vehicles involved) between the rural and total records.

Table 2 General information of the extracted datasets for rural routes.

Year	Total number of records	Total number of fatalities	Total number of injuries	Total number of pedal cyclists involved	Total number of pedestrians involved	Total number of motor vehicles involved	Proportion of the total number of records
2010	18,032	281	13,417	35	115	30,552	6.74%
2011	18,079	415	13,278	54	222	30,869	6.98%
2012	18,890	344	12,666	62	147	32,188	7.20%
2013	17,812	325	11,792	63	161	30,874	6.75%
2014	19,690	346	14,359	57	121	36,228	6.91%
2015	20,789	474	13,586	73	160	36,315	6.70%
2016	22,104	345	13,133	39	159	38,957	6.86%
Average	19,342	361	13,176	55	155	33,712	6.87%
Total	135,396	2,530	92,231	383	1,085	235,983	6.88%

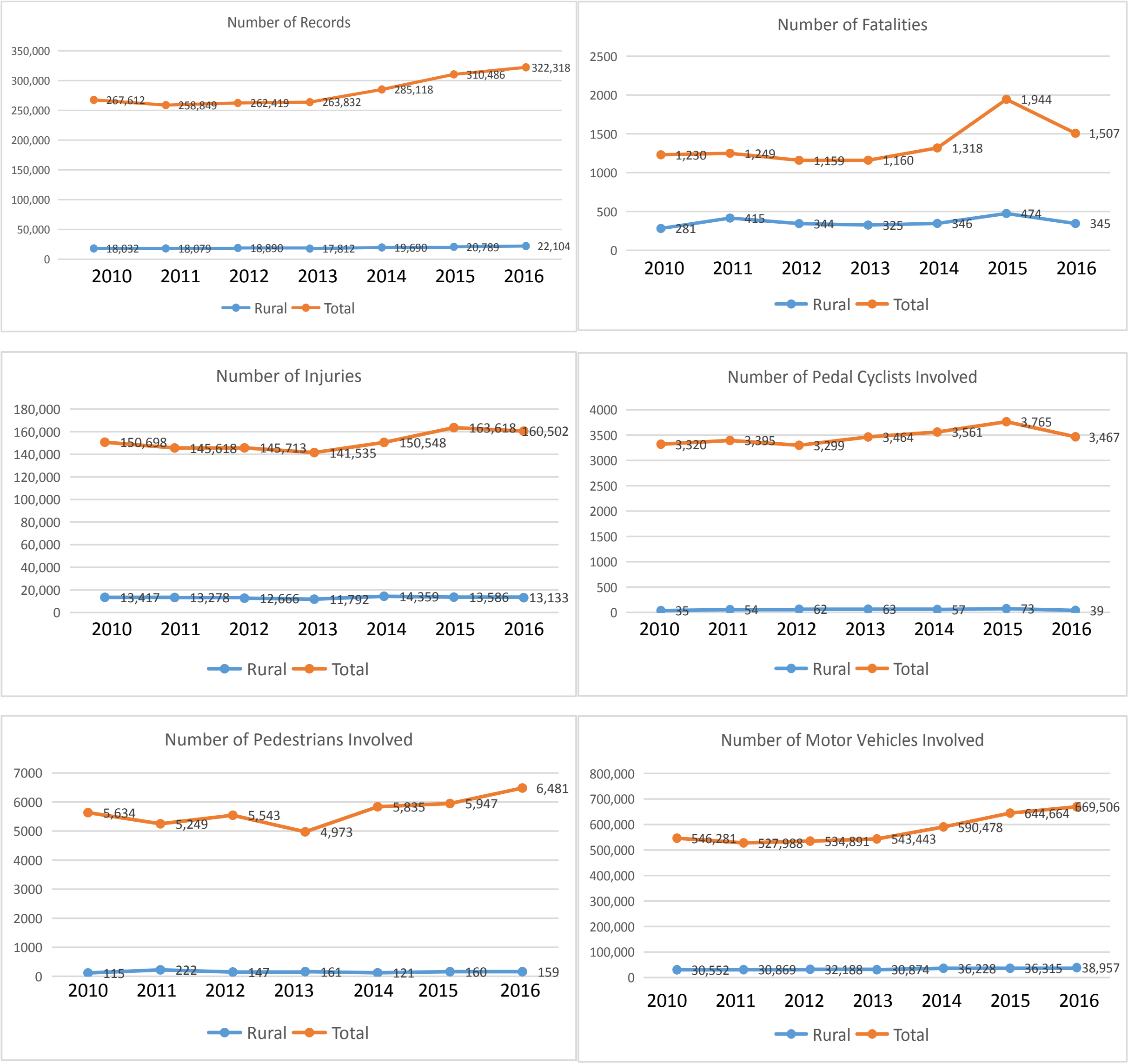


Figure 9 Dynamic trends of general information for rural and total records.

As shown in Figure 9, the data indicated that the dynamic trends of rural fatalities, injuries, pedal cyclists involved, pedestrians involved, and motor vehicles involved are different from those of the total records. For the total records, the number of collision records and the number of motor vehicles involved have a slightly increasing trend, while most of the other general information indices (i.e., number of fatalities, number of injuries, number of pedal cyclists involved, and number of pedestrians involved) fluctuated from 2010 to 2016, where the number of fatalities reached the highest value in 2015. For the rural records, the number of fatalities fluctuated, while all the other general information indices (i.e., number of collision records, number of injuries, number of pedal cyclists involved, and number of pedestrians involved) were somewhat stable. Comparison between the rural and total trend curves implies that the statistics of crashes in rural areas are different from those in the urban areas. Thus, we cannot simply apply the insights got from the urban areas to the rural ones.

3.1.2. Data from Tribal Communities

Since a lot of crashes occurred in the RITI communities are underreported, the reported crash data for rural routes may not be very accurate. However, the tribes in the rural areas have their own transportation safety datasets which could be a good complementary source for the WSDOT's crash datasets. There are twenty-nine federally-recognized tribes in Washington State. Native American reservations often include a mix of tribal, state, county, and city roads, which create jurisdictional complexities in terms of law enforcement, collision reporting, road maintenance, and capital safety projects. Through the Centennial Accord, the State of Washington and tribes have formally committed to working together on a government-to-government basis to address a number of common problems, including traffic safety issues. Having accurate data are essential to understanding these safety problems, selecting appropriate countermeasures, and evaluating performance. Without such supporting data, the evaluation, analysis, and diagnosis efforts of traffic safety become more difficult, if not impossible. It is also more difficult for RITI communities to compete for safety funding and justify their needs when lack of supporting data.

In order to collect the data from tribal communities, our research team interviewed or presented to 23 tribal leaders or transportation officials in Washington State through email, phone, and/or in person. We finally established positive connections with twelve tribes (i.e., Colville, Spokane, Muckleshoot, Swinomish, Yakama, Makah, Quinault, Skokomish, Puyallup, Lummi, Tulallp, and Sauk-Suiattle), including four strong connections (i.e., Colville, Spokane, Muckleshoot, and Swinomish) as shown in Figure 10. Ultimately, a formal research agreement with one tribe, i.e., Colville, was achieved. Through this agreement, Colville tribe agrees to provide us datasets which include fatal and serious injury crashes in Colville community over a period of 10 years. In order to protect the privacy of the tribe communities, the agreement requires that these datasets cannot be set as public source. However, this could be regarded as a major success as it lays the foundation for other tribes to share their crash data and allows the researcher to begin analysis work.

The crash data obtained from Colville tribe has been integrated with the WSDOT's crash data in the regional multi-source database system. However, due to the confidentiality requirement of the agreement with the Colville tribe, the details of the dataset are not described here.

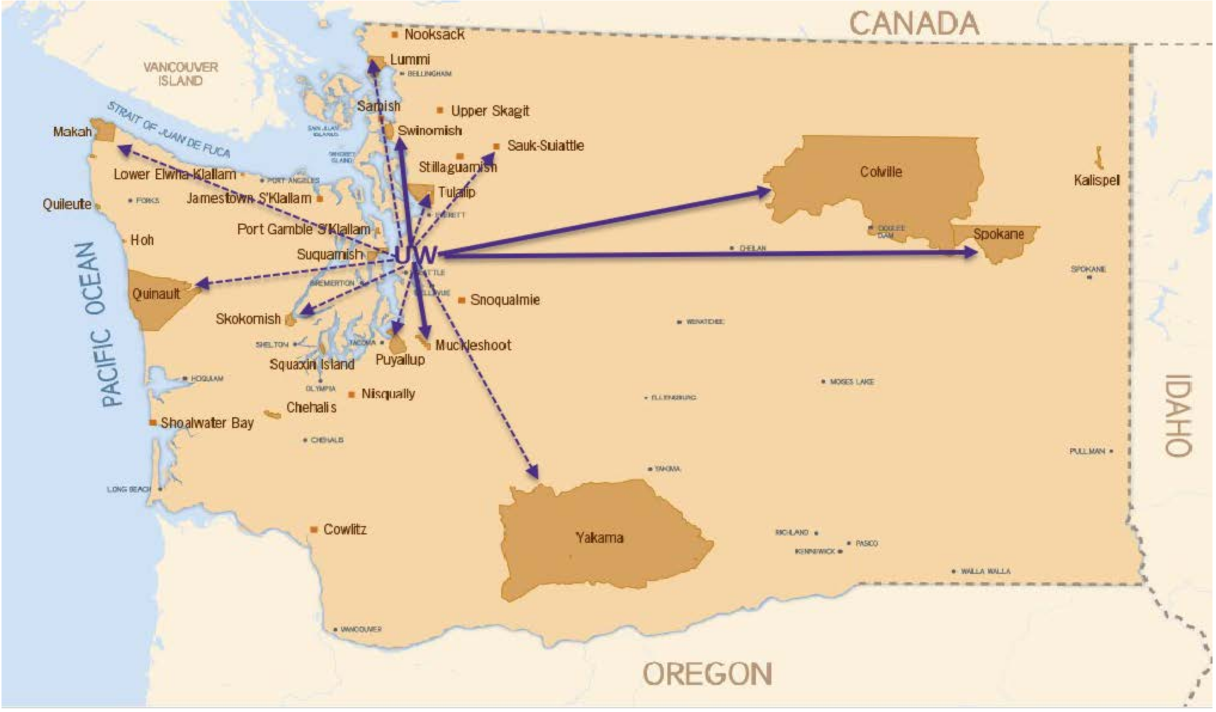


Figure 10 Established connections with tribal communities in Washington State.

3.2. Database System Architecture

The multi-source database fusion and integration system architecture was designed as shown in Figure 11. The data platform can be regarded as an extension of the urban road safety analysis module on DRIVE Net. New safety analysis functions were developed focusing on addressing the characteristics of the RITI communities in Washington State. The data platform system was developed based on state-wide multiple-year related safety data collected in Washington provided by WSDOT, including crash attributes and contribution factors, crash-related trauma data and medical records, traffic flow characteristics, roadway inventories, weather conditions, etc. Analytical methods were investigated to fuse and integrate multi-source data into one relational database. Microsoft SQL server 2012 was used to implement the database and manage the data. Entity-relational database diagram was designed and employed. Data consistency and integrity were checked and ensured. The relationships between these imported files were configured to make sure that they are connected correctly. The baseline data platform was developed on the web server as shown in Figure 11. The crash database was directly connected to the data visualization platform for crash data management. The online statistical data analysis functions were designed and implemented based on hotspot identification method to illustrate crash statistics in RITI communities in Washington State. For the crash data from the Colville tribe, a separate rural safety analysis module was developed on the database system. A password is needed for users to access this module. Only those authorized to use the data by the UW-Colville agreement have the password.

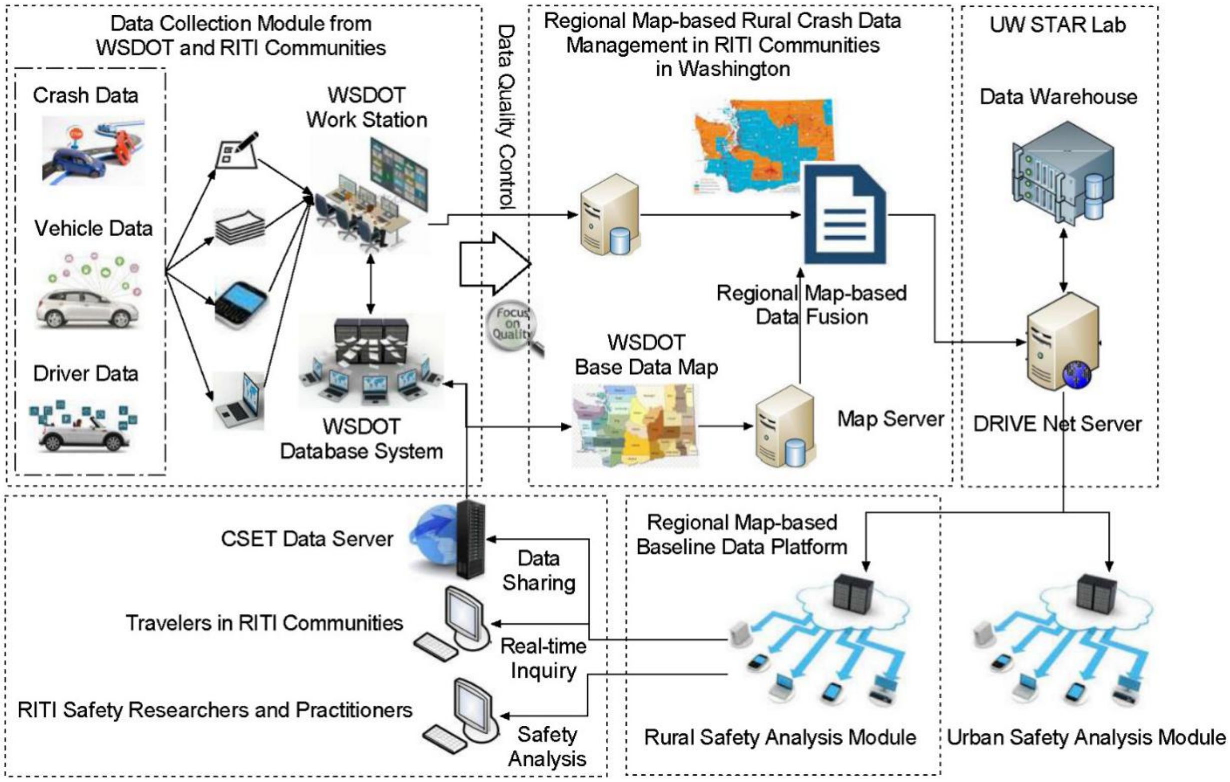


Figure 11 Regional Map-Based Baseline Data Platform System Architecture.

CHAPTER 4. SAFETY DATA MANAGEMENT AND ANALYSIS

One of the purposes of this project is to identify and document existing multi-source crash data sources in rural, isolated, tribal, and indigenous (RITI) communities in Washington State. The multi-source data including traffic flow characteristics, crash attributes and contributing factors, crash-related trauma data and medical records, weather conditions, etc., were collected from WSDOT, the Planning Department of Tribes in rural and isolated communities, and other data providers in Washington State. This chapter will discuss the data storage and management as well as data visualization and analysis.

4.1. Data Storage and Management

4.1.1. Data Storage

All of the data were managed and stored in the SQL server database format. The existing DRIVE Net server at the UW STAR Lab was used to host the datasets. Potential users of the data might include but are not limited to researchers, scientists, engineers, students, and transportation agencies. As part of establishing baseline crash data infrastructure, the proposed data platform can enable effective traffic safety program management at all levels in RITI communities, and it can further help with design and implementation of appropriate countermeasures to mitigate rural crash risks and reduce associated crash severities.

4.1.2. Data Standards and Quality Control

Safety-related data were collected and stored in a text file format by WSDOT and the Colville Tribal Transportation Planning Department, though the tribal data are kept separate so as not to be included in visualization tools that were made publicly available. User query files and system configuration files were saved and stored in the text files and shared among authorized users. These raw data were transferred into the database format for improving data quality and consistency. Microsoft SQL Server 2012 was used to implement the database and manage the data. The data quality control process was conducted to guarantee timeliness, accuracy, completeness, and consistency of the data; an example operation includes correcting location information for crashes.

The importance of data quality control is widely recognized by most transportation agencies. There are three principles involved in the data quality control process. First, a rule-based error detection algorithm that uses criteria developed to address common data errors is applied to flag erroneous and questionable observations. Second, a sensitivity adjustment algorithm is applied to detect data 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. Metadata for spatial safety-related records used in this effort was maintained in the standard ArcGIS format. All XML-based and binary formats used for storing data were well-established and documented. Based on the above three principles, a six-step method was employed to clean the data by wiping out the outliers from spatial and temporal aspects as shown in Figure 12.

Step 1: Feature Calculation. We have defined corresponding features for each collision record to characterize the spatiotemporal relationship between sampling points. That includes: crash frequency, locations, severities, roadway segment length, average number of lanes (NOL), horizontal curve type

(HCT), curvature of the segment (COS), average width of outer shoulder (WOS), average width of inner shoulder (WIS), average width of median (WM), dominant lane surface type (DLST), dominant outer shoulder type (DOST), dominant inner shoulder type (DIST), dominant median type (DMT), average speed limit (ASL), annual average daily traffic (AADT), AADT per lane, road surface conditions (RSC, i.e., dry, wet, snow/ice/slush), and visibility (good, bad). Anomalies on these features may indicate quality problems. For example, small number of sampling points suggests data missing; large max segment speed suggests jumping points; large sum of turning angle suggests signal oscillation. The first principle, rule-based error detection algorithm, is employed in this step to identify the outliers or clusters.

Step 2: Distribution Visualization. We assume a quality problem corresponds to some attributes of a record outliers or clusters. We try to visualize them in several views, including: a temporal histogram showing the temporal distribution, a map showing the spatial distribution, a high dimensional projection view and a parallel coordinates view showing the feature space distribution. The colors of the attributes of the records in all views indicates the corresponding types of quality problems. The second principle, sensitivity adjustment algorithm, is employed in this step to adjust the outliers or clusters.

Step 3: Interactive Filtering. We extract the outlier or clustered of the attributes of the records with a set of filters. We select a time range in the temporal histogram view, a rectangular spatial range in the map, a rectangular range in the projected feature space, and value ranges on parallel coordinates axes. The attributes of the records satisfying all filters will be selected. We do not automatically generate clusters, because the result directly generated by machine may not correspond to quality problems, and can be hard to interpret. We rely on human analysts to find meaningful clusters.

Step 4: Detail Visualization. All filtered attributes of the records are maintained in a list. We can highlight one and get its detailed information with three visualizations. This helps to determine whether the attributes of the records are normal or having quality problems. The map view now shows their path and shape. A timeline was created to show the temporal change of attributes such as turning angle, segment time interval, segment distance and segment average speed. For each attribute we estimated a valid value range, and mark the range yellow. Values outside the yellow range can be considered as potential outliers.

Step 5: Known quality problem detection and separation. For each quality problem, a binary support vector machine (SVM) classifier is automatically built. As manual labelling is very expensive, we use active learning strategy to mitigate the labeling effort. That is, the system will ask us to label the attributes of the records that have greater influence on the model accuracy. We can also search for attributes of the records similar to labelled ones, and label them. These together help to extract attributes of the records of identified quality problems in a semi-automatic way. We can separate out these attributes of the records and focus on the remaining ones. When the remaining ones are all normal, our discovery process ends.

Step 6: Data imputation. The imputation algorithm is applied to fill in missing observations and estimate prediction intervals. Random imputation method is adopted to randomly generate the data within the reasonable ranges.

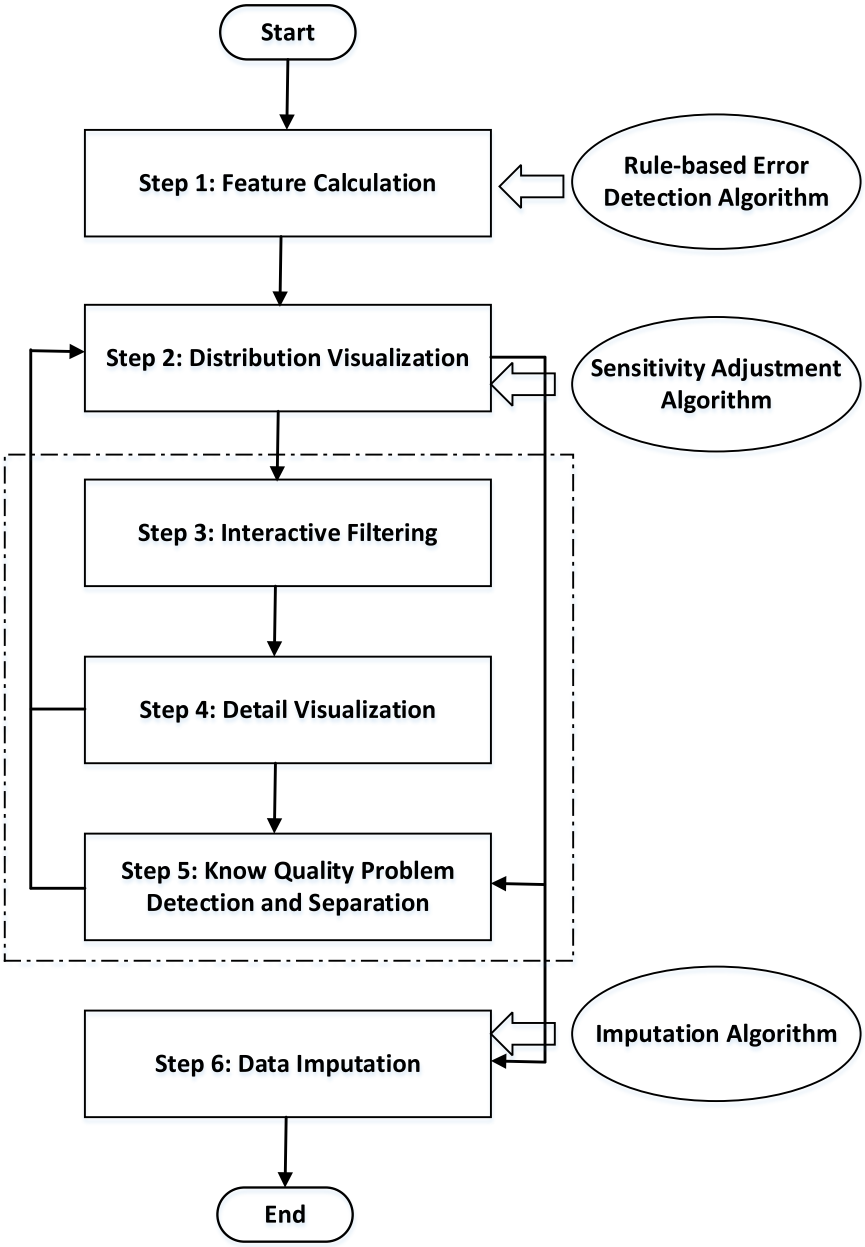

Start
Step 1: Feature Calculation
Rule-based Error Detection Algorithm
Step 2: Distribution Visualization
Sensitivity Adjustment Algorithm
Step 3: Interactive Filtering
Step 4: Detail Visualization
Step 5: Know Quality Problem Detection and Separation
Imputation Algorithm
Step 6: Data Imputation
End

Figure 12 Flowchart of data quality control process.

4.1.3. Data Access Policies

The data access policies were set up by strictly following the CSET data management plan. The multi-source data is shared through local computer server stations. The tribal crash data was made accessible to authorized users so they can download the datasets with a username and password; currently only the researchers and Colville Tribe have access to the data, though Colville Tribe has made some of their data available in an online map through funding from Washington Traffic Safety Commission. The WSDOT crash data were set to be public for all users. The research data and information sharing activities are monitored by the project team as well as CSET administrative staff. In order to protect privacy, confidentiality, security, intellectual property, and/or other rights, public access is regulated. Privileged or confidential information will be released only in a form that protects the privacy of individuals and subjects involved. Whenever possible, data conflicting with privacy requirements would be made anonymous to enable sharing it with the public. Before data is stored, it is stripped of all institutional and individual identifiers to ensure confidentiality following procedures developed by the researchers. With respect to the public, some parts of the data will be made available after researchers have successfully published the main findings in peer reviewed journals and after working with the necessary DOT personnel. The data visualization system produced by the WSDOT crash data is accessible by the public. The data servers will allow direct electronic transfer of data layers and access to supporting documentation by authorized parties at any time.

4.1.4. Data Archiving and Preservation

The server station is password-protected in a safe and effective manner, with sufficient provisions for backup and recovery in case of equipment failure. To meet the technical specifications of a trusted digital repository, the server room is built to conform to the standard reference models and includes a framework of policies and procedures that enable access to original datasets. In order to ensure long-term archiving and preservation, technical design considerations capture technical and descriptive metadata to permit long-term preservation and collection management, verification of dataset integrity, and adequate backup and mirroring procedures. Digital Object Identifiers (DOIs) were attached to all data stored from this project.

4.2. Data Visualization and Analysis

Visualization and visual analysis are important for a highly efficient data-driven analytical platform. In this project, a safety analysis module was developed for visualizing the data in the regional multi-source database system in RITI communities. This is an importation component of the system because it allows for easier interpretation of the data and analysis results and can be used in report generation. The visualization tools are meant to be user-friendly and intuitive, at least for the basic use-cases. The visualization functionality and analytical functionality are introduced as below.

4.2.1. Visualization Functionality

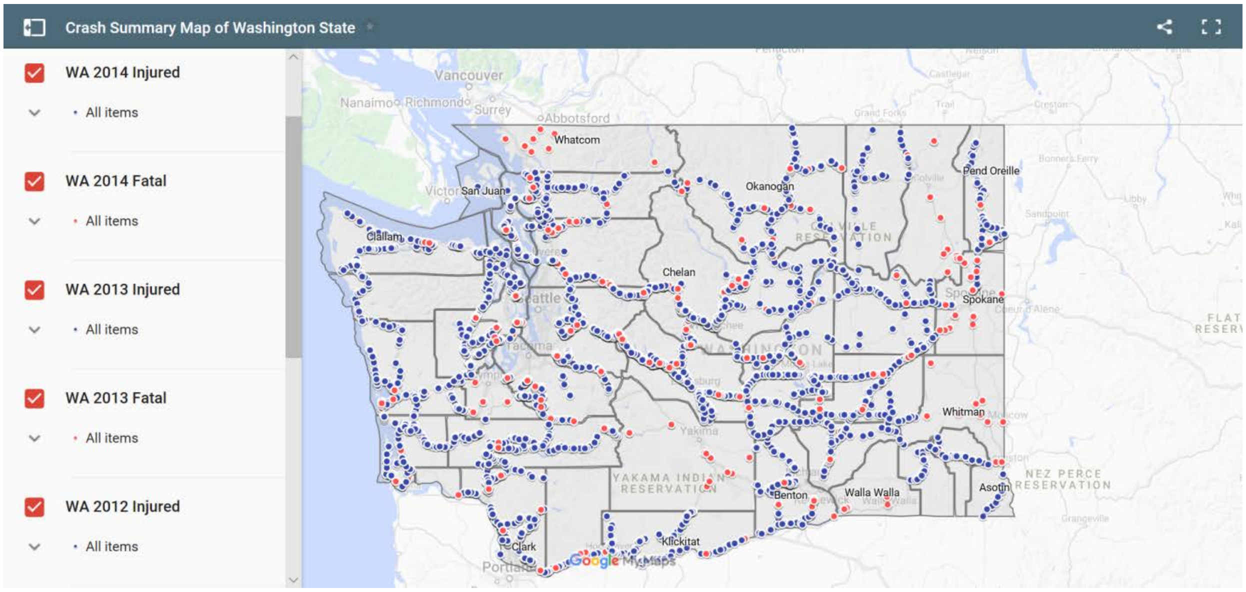
The data visualization platform is developed based on the Vaadin Framework. Vaadin is an open-source platform for web application development. The Vaadin platform includes a set of web components, a Java web framework, and a set of tools and application starters. Figure 13 shows a snapshot of the interface of the developed data visualization platform, which was created using Google Maps features and a Wix landing page. The Vaadin components can be utilized to expand this demo or create an entirely new interface, either using a Google Maps API or by using the demo to simply validate the functionality of the Vaadin application. Google Maps and Wix were suitable for the demo version of the data visualization platform due to the low crash data volume and the familiar interface. The extracted crash data on the rural routes in Washington State from WSDOT was employed to do the data visualization demo. Future functionality would ensure robustness for handling larger amounts of data.

Crash Summary Map of Washington State

IN THE RURAL AREA

Figure 13 Interface of the safety module for RITI communities in the database system.

As shown in Figure 14, each point represents a collision/crash record on the rural routes of Washington State. A red point indicates a fatal crash while a blue one denotes an injury crash. From the options in the left window, the users can select specific years or crash severities to show the locations of those crashes in a Google Map-based interface. In order to identify the location of each crash on the map, the attribute "(state plane x, state plane y)" for each crash record in the dataset was employed to be transformed into longitude and latitude values which can be used to locate the position for each crash. This visualization function could be helpful to present the spatial and temporal distribution of different types of crashes (i.e., fatalities and injuries), which lays the foundation for developing hotspot identification analytical functions in this database system.



Crash Summary Map of Washington State

☑ WA 2014 Injured
- All items

☑ WA 2014 Fatal
- All items

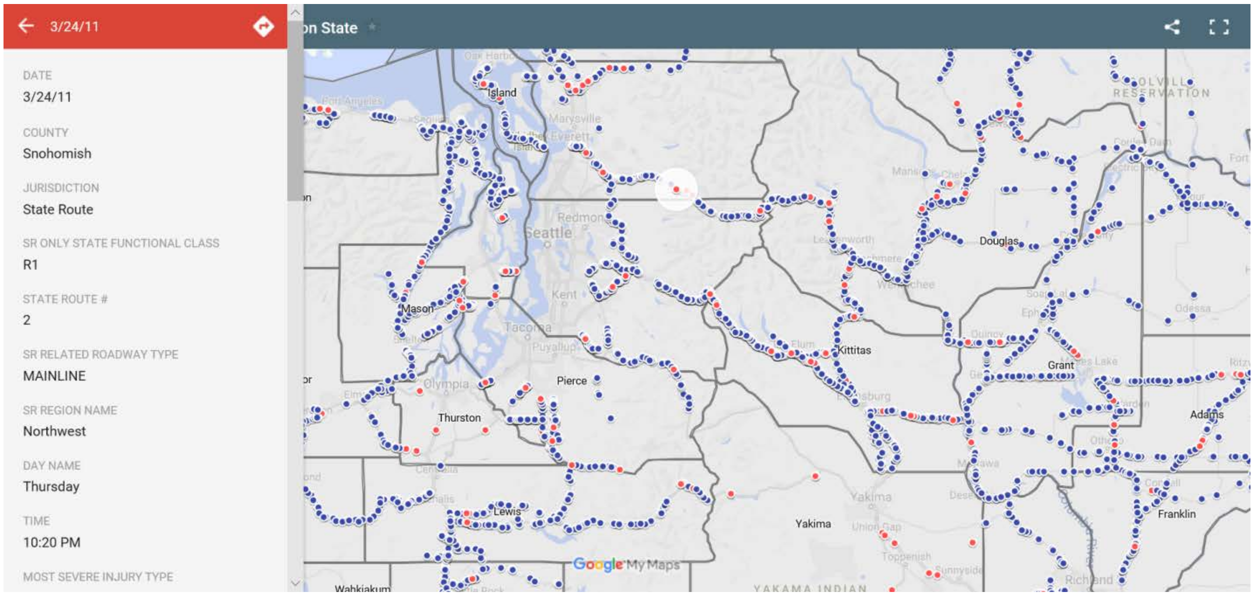
☑ WA 2013 Injured
- All items

☑ WA 2013 Fatal
- All items

☑ WA 2012 Injured
- All items

Figure 14 Year and crash severity options in the data visualization platform.

Furthermore, as shown in Figure 15, the users can also interact with the crashes by selecting a specific point to see its detailed information such as collision report date, county, route type, route direction, day, time, and so on. Importantly, this interaction can also show important crash characteristics such as if impairment was involved, weather conditions, and contributing circumstances listed in the police report. These, among other factors, are all very important for rural area analysis because behavioral-influenced crashes occur at a higher rate in rural areas than in urban areas. However, only reviewing the basic information for each crash may not be enough to properly identify the correlations among different crashes, thus, an interactive analytical function was developed to analyze a group of crashes as illustrated in Figure 16. Also, it is important to note that all the information available to view for each crash is still linked to the SQL Server database, so that greater analysis on these data can still be conducted, this demo is just not able to visualize more complex analysis and results. Still, this demo serves the purpose of this project well enough and the user interface is certainly similar to what is expected.


3/24/11

DATE
3/24/11

COUNTY
Snohomish

JURISDICTION
State Route

SR ONLY STATE FUNCTIONAL CLASS
R1

STATE ROUTE #
2

SR RELATED ROADWAY TYPE
MAINLINE

SR REGION NAME
Northwest

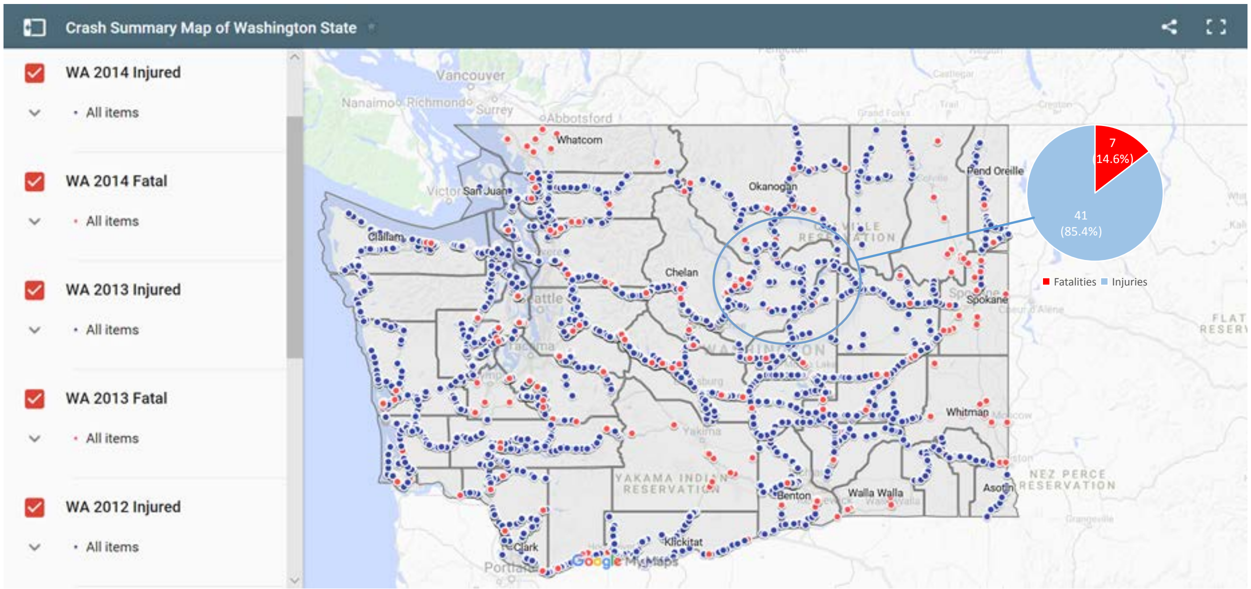
DAY NAME
Thursday

TIME
10:20 PM

MOST SEVERE INJURY TYPE

Figure 15 Detailed information for a specific crash record in the data visualization platform.

The users can interact with the interface by creating a circle to select a group of crashes. The statistics of the group of crashes will be displayed such as the frequency and proportion of each crash type. For example, in Figure 16, the interface shows that there are 7 (14.6%) fatal crashes and 41 (85.4%) injury crashes in the created circle. This is perhaps a user-friendly feature in terms of a simple action being able to produce greater crash information, but may not be the way we incorporate such a feature in the future because the shape cannot be easily controlled.



Crash Summary Map of Washington State

☑ WA 2014 Injured
- All items

☑ WA 2014 Fatal
- All items

☑ WA 2013 Injured
- All items

☑ WA 2013 Fatal
- All items

☑ WA 2012 Injured
- All items

Figure 16 Interactive analytical functions for a group of crash records in the data visualization platform.

4.2.2. Analytical Functionality

Hotspot identification (HSID) for the high-risk roadway segments is of great importance to assisting transportation and government authorities in their efforts to improve the safety of roadways. It was indicated that the crash severity should not be neglected in the HSID process. The hotspots corresponding to high crash risk locations can be quite different when considering the crash frequency by different levels of crash severity. It is particularly important to take into account crash severities in site ranking, because the cost of crashes could be significantly different at different severity levels. For instance, a road segment with a higher frequency of fatal accidents may be considered more hazardous than a road segment with fewer fatal incidents, but with more serious or minor injury incidents. As a result, a risk weight factor was developed in this database system by combining the average crash cost with the corresponding probability for each type of crash severity. A new safety performance index (SPI) and a new potential safety improvement index (PSII) were also developed by using the risk weight and the generalized nonlinear model-based mixed multinomial logit ((GNM-based MNL) approach (Zeng et al., 2017) combined with the traditional Empirical Bayes (EB) method.

In the HSID demo using this database system, the crash data for RITI communities were aggregated on the roadway segments and divided into three categories, i.e., property damage only (PDO) (*k*=1), injury (*k*=2), and fatal (*k*=3). To illustrate the mechanism of the HSID algorithm, we have the following definitions. The crash severity type, denoted by *Y*, was defined as the response variable, whereas contributing factors such as roadway geometric characteristics, traffic characteristics, and weather conditions were the independent variables denoted by x_{ij} , where *i* denotes the roadway segment and *j* is the index for the independent variables. *Y* is expressed as below:

$$Y = \begin{cases} 1, \text{ if crash is PDO,} \\ 2, \text{ if crash is injury,} \\ 3, \text{ if crash if fatal,} \end{cases} \qquad (4.1)$$

In order to account for the unobserved heterogeneity, let Ω=$(\omega_1, \omega_2, \omega_3)$, and note that the Ω vector has a continuous density function $f(\Omega | \Gamma)$, where Γ is a vector of parameters charactering the density function, $\omega_k = [\omega_{k1}, \omega_{k2}, \ldots, \omega_{kJ}]^T$ is the coefficient vector for the kth category of the predictor vector. Based on the GNM-based MNL model (Zeng et al., 2017), the expected crash density for different severity levels can be calculated as follows:

(1) Expected PDO crash density:

$$d_{i1} = \int e^{U_i \omega + \beta_0} f(\omega | \varphi) d\omega \cdot \int \frac{e^{U_{1i}\omega_1 + \beta_{10}}}{1 + \sum_{k=1}^{2} e^{U_{ki}\omega_k + \beta_{k0}}} f(\Omega | \Gamma) d\Omega, \; i = 1, 2, \ldots, n, \qquad (4.2)$$

(2) Expected injury crash density:

$$d_{i2} = \int e^{U_i \omega + \beta_0} f(\omega | \varphi) d\omega \cdot \int \frac{e^{U_{2i}\omega_2 + \beta_{20}}}{1 + \sum_{k=1}^{2} e^{U_{ki}\omega_k + \beta_{k0}}} f(\Omega | \Gamma) d\Omega, \; i = 1, 2, \ldots, n, \qquad (4.3)$$

(3) Expected fatal crash density:

$$d_{i3} = \int e^{U_i \omega + \beta_0} f(\omega | \varphi) d\omega \cdot \int \frac{1}{1 + \sum_{k=1}^{2} e^{U_{ki}\omega_k + \beta_{k0}}} f(\Omega | \Gamma) d\Omega, \; i = 1, 2, \ldots, n, \qquad (4.3)$$

where $U_{ki} = [U_{ki1}(x_{i1}), U_{ki2}(x_{i2}), \ldots, U_{kiJ}(x_{iJ})]$ is the nonlinear predictor vector of roadway segment *i* for contributing factors; β_{k0} is an intercept term specific to crash severity type *k*.

To develop the new indices SPI and PSII, the equivalent property damage only (EPDO) method (Washington et al., 2014) is modified and employed to weight crashes according to severity to develop a combined crash density and severity score (CCDSS) for each roadway segment. The weight factors are based on PDO crash costs. An EPDO value summarizes the crash costs and severity. In the calculations, weight factors were assessed from the crash cost estimates developed by WSDOT in the Annual Collision Data Summary Reports (2011-2014). Using average crash costs for motorways, fatal crashes ($2,227,851) have a weight factor equal to 981, injury crashes ($20,439) have a weight factor equal to 9, and PDO crashes ($2,271) have a weight factor equal to 1. In contrast to the traditional method, a new risk weight factor is defined as below:

$$F_w = \frac{c_F \cdot \eta_F}{c_P \cdot \eta_P}, \; I_w = \frac{c_I \cdot \eta_I}{c_P \cdot \eta_P}, \; P_w = 1, \qquad (4.4)$$

Where F_w, I_w, and P_w denote the risk weigh factor for fatality, injury, and PDO respectively; $c_F = \$2,227,851$, $c_I = \$20,439$, and $c_P = \$2,271$ are the average costs for fatal, injury, and PDO crashes; η_F, η_I, and η_P are the probabilities of occurrence for fatal, injury, and PDO crashes.

Then, the expected CCDSS (ECCDSS) for roadway segment i can be defined as:

$$ECCDSS_i = d_{i1}F_w + d_{i2}I_w + d_{i3}P_w,\ i = 1, 2, \ldots, n, \tag{4.5}$$

Based on the EB method, the SPI can be defined as below:

$$SPI_i = \lambda_i ECCDSS_i + (1-\lambda_i)OCCDSS_i,\ i = 1, 2, \ldots, n, \tag{4.6}$$

where $OCCDSS_i$ is the observed combined crash density and severity score (OCCDSS) for roadway segment i and is defined as:

$$OCCDSS_i = \sigma_{i1}F_w + \sigma_{i2}I_w + \sigma_{i3}P_w,\ i = 1, 2, \ldots, n, \tag{4.7}$$

where σ_{i1}, σ_{i2}, and σ_{i3} are the observed fatal, injury, and PDO crash density along segment i during a certain year respectively; λ_i is a weighting factor that is calculated through the following equation:

$$\lambda_i = \frac{1}{1+\alpha_i ECCDSS_i} \tag{4.8}$$

where α_i is the overdispersion parameter, which is a constant for a given model and is derived during the regression calibration process.

The PSII was developed as the difference between the SPI can the ECCDSS, as follows:

$$PSII_i = SPI_i - ECCDSS_i = (1-\lambda_i)(OCCDSS_i - ECCDSS_i),\ i = 1, 2, \ldots, n,$$

(4.9)

If the observed combined crash density and severity score is greater than the expected one, PSII will be positive, which means there is potential safety improvement existed. Otherwise, if the observed combined crash density and severity score is smaller than the expected one, PSII will be negative, which implies that the potential for safety improvement is not meaningful. Greater attention and resources should be focused on the areas with higher potential safety improvements.

The SPI developed in the preceding is used to color-code the regional map based on safety performance. The PSII is employed to highlight potential safety improvements on the map. By combining the two indices on the regional map, one can easily identify crash hotspots and the key influencing factors to consider in an improvement package.

4.2.3. Comparison Tests of Indexes

In order to test the effectiveness of using the developed indexes SPI and PSII for hotspots identification (HSID), four reference performance indexes were introduced, including the expected crash density based on the conventional safety performance function from the Highway Safety Manual (HSM), i.e., the Negative Binomial Generalized Linear Model (NB GLM), expected crash density based on the Generalized Nonlinear Model (GNM), Empirical Bayes (EB) estimated crash density based on the NB GLM, and EB estimated crash density based on the GNM.

In this study, the extracted crash data on the rural state routes from 2010 to 2016 from WSDOT were employed for the comparison tests. The total number of crashes during the period on the rural state routes was 135,396, including 2530 fatalities, 92,231 injuries, and 40,635 PDO crashes. Thus, the probabilities of occurrence for fatal, injury, and PDO crashes are $\eta_F = 0.0187$, $\eta_I = 0.6812$, and $\eta_P = 0.3001$. According to Eq. (4.4), the values of the risk weight factors are $F_w = 61.129$, $I_w = 20.429$, and $P_w = 1$.

Two evaluation test methods for HSID developed by Cheng and Washington (2008) were employed to do the comparison test, including the site consistency test, method consistency test, total rank differences test, and the total score test. The evaluation uses the following general procedure:

(1) For comparing the four reference performance indexes with SPI and PSII, the 7-year data were separated into two periods, Period 1 (Year 2010-2013) and Period 2 (Year 2014-2016).

(2) Road segments are sorted in descending order based on the six indexes according to their corresponding criteria.

(3) Road segments with the highest rankings are flagged as hotspots. Typically, a threshold is assigned according to safety funds available for improvement, such as the top 1% of sites. In this study, both the top 1% and 5% of the sites are used.

4.2.3.1 Site Consistency Test

The site consistency test (SCT) measures the ability of an index to consistently identify a high-risk site over repeated observation periods. There is a basic assumption of the test that a road segment identified as high risk during time period i should also reveal an inferior safety performance in a subsequent time period t+1, given that the road segment is in fact high risk and no significant changes have occurred at the road segment. The index that identifies road segment in a future period with the highest crash frequency is the most consistent. In this study, the SPI developed above is employed as the safety performance criterion in the subsequent time period. The test statistic is expressed as below:

$$SCT_{h,t+1} = \sum_{q=n-n\cdot\gamma+1}^{n} HSID_{q,h,t+1},\ h = 1, 2, \mathrm{K}\ , H, \qquad (4.10)$$

where *h* indicates the index being compared; *n* is the total number of road segment, *γ* is the threshold of identified hotspots (e.g., *γ* = 0.01 corresponds with top 1% of *n* road segments identified as hotspots, and *n* ·*γ* is the number of identified hotspots); $HSID_{q,h,t+1}$ is the value of *h*th index for road segment *q* at time period *t*+1.

Table 3 shows the results of site consistency test of the six indexes for HSID. The results indicate that the SPI outperforms other indexes in identifying both of the top 1% and 5% of hotspots with highest SCT values, 324,536.23 and 1,802,411.83, in Period 2, followed closely by the EB Crash Density (GNM) index.

Table 3 Results of site consistency test of various indexes for HSID.

No.	Index Name	$SCT_{h,t}$			
		γ=0.01		γ=0.05	
		Period 1	Period 2	Period 1	Period 2
1	SPI	444,373.01	324,536.23	2,271,344.89	1,802,411.83
2	PSII	443,872.64	323,701.04	2,270,982.74	1,802,112.28
3	Crash Density (NB GLM)	443,011.97	323,318.93	2,270,310.04	1,801,538.12
4	Crash Density (GNM)	443,581.93	324,024.18	2,270,675.52	1,801,700.43
5	EB Crash Density (NB GLM)	443,326.88	323,913.46	2,270,803.01	1,801,943.55
6	EB Crash Density (GNM)	444,265.24	324,317.34	2,271,061.27	1,802,203.14

4.2.3.2 Method Consistency Test

The method consistency test (MCT) evaluates an index performance by measuring the number of the same hotspots identified in both time periods. The assumption of this method requires that road segments are in the same or similar underlying operational state and their expected safety performance remains virtually unaltered over the two analysis periods. With this assumption of homogeneity, the greater the number of hotspots identified in both periods the more consistent the performance of the HSID method. The test statistic is given as below:

$$MCT_h = \{s_{n-n\cdot\gamma+1}, s_{n-n\cdot\gamma}, \ldots, s_n\}_{h,t} \cap \{s_{n-n\cdot\gamma+1}, s_{n-n\cdot\gamma}, \ldots, s_n\}_{h,t+1},\ h = 1, 2, \ldots, H, \quad (4.11)$$

where, only segments $\{s_{n-n\cdot\gamma+1}, s_{n-n\cdot\gamma}, \ldots, s_n\}$ identified in the top threshold γ are compared. Table 4 shows the number of similarly identified hotspots by the six indexes over the two periods.

The results indicate that the SPI is superior in this test by identifying the largest number of the same hotspots in both cases of γ=0.01 and γ=0.05, with 311 and 1775 road segments, respectively. In other words, the SPI method identified 311 segments in Period 1 that were also identified as hotspots in Period 2. The EB Crash Density (GNM), which performs slightly better than the EB Crash Density (NB GLM), places 2nd with identifying 277 consistent hotspots (in the case of γ=0.01) and 1685 consistent hotspots (in the case of γ=0.05). Table 4 also shows the percentage of the identified same hotspots for the six indexes. It can be found that there is a consistent drop in percentages as threshold value drop. The reason is that the top segments suffer from greater random fluctuations in crashes, and thus the higher is the threshold, the larger are the random fluctuations and the likelihood of not being identified in a prior period.

Table 4 Results of method consistency test of various indexes for HSID.

No.	Index Name	$MCT_{h,t}$	
		γ=0.01	γ=0.05
1	SPI	311 (53.6%)	1775 (61.2%)
2	PSII	257 (44.6%)	1471 (50.7%)
3	Crash Density (NB GLM)	230 (39.6%)	1314 (45.3%)
4	Crash Density (GNM)	241 (41.5%)	1439 (49.6%)
5	EB Crash Density (NB GLM)	251 (43.3%)	1572 (54.2%)
6	EB Crash Density (GNM)	277 (47.8%)	1685 (58.1%)

Overall, the two tests reveal that the SPI is the most consistent and reliable index for HSID. Although it can only be applied to road segments where the crash data for different severities are available, with the rapid development of intelligent transportation systems and data collection technologies, this index (i.e., SPI) could become quite useful in identifying high-risk road segments.

CHAPTER 5. CONCLUSIONS

This project aims to develop a regional multi-source database system of traffic safety data management and analysis for RITI communities in Washington State. As part of baseline crash data infrastructure establishment, the proposed data platform can enable effective traffic safety program management at all levels in RITI communities, and design and implement appropriate countermeasures to mitigate rural crash severities and risks.

In order to understand the current state-of-the-art and practice, the existing related web-based archived data user service systems were investigated and classified into three categories: performance measure system, multiple-agency system, and online data-driven analysis system. The research team found that various web-based archived data user service systems have been developed during the past decade. However, most of them focus on travel time analysis, and little existing effort has been concentrating on multi-source safety data fusion and integration for RITI communities.

For identifying and documenting existing crash data sources in RITI communities in Washington, a two-fold strategy was implemented to collect the data from multiple sources: (1) Collect raw data from WSDOT and extract the related crash data in the rural areas; (2) Collect safety data from tribal communities in Washington State. WSDOT provided crash data on the state routes from 2010 to 2016. The datasets include 266 attributes for each record including collision report number, state route type (urban/rural), number of fatalities, number of injuries, number of pedal cyclists involved, number of pedestrians involved, and number of motor vehicles involved. The crash data on rural routes were extracted from the raw data provided by WSDOT and integrated into the multi-source database system, including traffic flow characteristics, crash attributes and contribution factors, crash-related trauma data and medical records, and weather conditions. In order to collect the data from tribal communities, our research team interviewed or reached out by other means to 23 tribal leaders or transportation officials in Washington State through email, phone, and/or in person. The team established positive connections with twelve tribes (i.e., Colville, Spokane, Muckleshoot, Swinomish, Yakama, Makah, Quinault, Skokomish, Puyallup, Lummi, Tulallp, and Sauk-Suiattle), including four strong connections (i.e., Colville, Spokane, Muckleshoot, and Swinomish). Ultimately, a formal research agreement with one tribe, the Colville, was achieved. The Colville tribe provided the crash data in their tribal communities over a period of ten year under a confidentiality agreement.

A multi-source database fusion and integration system architecture was designed. The data platform can be regarded as an extension of the urban road safety analysis module on DRIVE Net. New safety analysis functions were developed focusing on addressing the characteristics of the RITI communities in Washington State. Microsoft SQL Server 2012 was used to implement the database and manage the data. The data quality control process was conducted to guarantee timeliness, accuracy, completeness, and consistency of the data. A six-step data quality control method was employed to clean the data by wiping out the outliers from spatial and temporal aspects. The data access policies were set up by strictly following the CSET data management plan. The multi-source data is shared through local computer server stations. The tribal crash data is accessible only to authorized users by the UW-Colville agreement so they can download the datasets by using password, while other datasets are accessible by all users. Digital Object Identifiers (DOIs) were attached to all data stored from this project.

A safety analysis module was developed for analyzing and visualizing the data in the regional multi-source database system for RITI communities. The data visualization platform is developed based on the Vaadin Framework. This visualization function could be helpful to present the spatial and temporal distribution of different types of crashes (i.e., fatalities and injuries), which lays the foundation for developing hotspot identification analytical functions in this database system. The users can also interact with the interface for data analysis. A safety performance index (SPI) and a potential safety improvement index (PSII) were also developed for crash analysis and hotspot identification which have a potential to become criteria for evaluating the high-risk roadway segments. The site consistency test and method consistency test were employed to compare the SPI and PSII with four other indexes for HSID. The test results demonstrated that the SPI is the most consistent and reliable index for HSID.

In summary, this research effort investigated existing data resources for RITI community safety equity improvement and build an online multi-source database system for safety data management, analysis, and visualization. A safety performance index and a potential safety improvement index were also developed and tested. These prototype tools and methods are valuable first step to identifying safety challenges and improvement directions for RITI communities.

REFERENCES

Bertini, R.L., Hansen, S., Byrd, A., & Yin, T., 2005. *Experience implementing a user service for archived intelligent transportation system data*. Journal of the Transportation Research Board, 1917, 90-99.

Chen, C., 2003. *Freeway Performance Measurement System (PeMS)*. Publication UCB-ITS-PRR-2003-22. California Partners for Advanced Transit and Highways (PATH), University of California, Berkeley, Richmond, Calif.

Cheng, W., Washington, S., 2008. New criteria for evaluating methods of identifying hot spots. Journal of the Transportation Research Board, 2083, 76-85.

Federal Highway Administration, 2012. *2012 Traffic Safety Facts “Rural and Urban Comparison”*. U.S. Department of Transportation.

Federal Highway Administration, 2014. *2014 Traffic Safety Facts “Rural and Urban Comparison”*. U.S. Department of Transportation.

Federal Highway Administration, 2016. *Local and Rural Road Safety Program*. U.S. Department of Transportation. https://safety.fhwa.dot.gov/local_rural/

Filippova, D., & Pack, M.L., 2009. *Mining multivariate accident data*. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

Lund, A.S., & Pack, M.L., 2010. *Dynamic wide-area congestion and incident monitoring using probe data*. Journal of the Transportation Research Board, 2174, 1-9.

Ma, X.L., Wu, Y.J., & Wang, Y.H., 2011a. *DRIVE Net e-science transportation platform for data sharing, visualization, modeling, and analysis*. Journal of the Transportation Research Board, 2215, 37-49.

Ma, X.L., McCormack, E.D., & Wang, Y.H., 2011b. *Processing commercial GPS data to develop a web-based truck performance measure program*. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C.

Ma, W., 2008. *A Real-Time Performance Measurement System for Arterial Traffic Signals*. PhD Dissertation, University of Minnesota.

National Association of Counties, 2009. *Rural Road Safety: Needs Assessment Summary Report*. https://www.naco.org/sites/default/files/documents/Rural-Roads%20Safety%20Fact%20Sheet.pdf

Pack, M.L., & Ivanov, N., 2008. *Web-based, interactive temporal-spatial traffic and incident data visualization tool*. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C.

Pack, M.L., Bryan, J.R., & Steffes, A., 2008. *Overview and status of the Regional Integrated Transportation Information System in the national capital region*. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C.

Pack, M.L., Weisberg, P., & Bista, S., 2005. *Four-dimensional interactive visualization system for transportation management and traveler information*. Journal of the Transportation Research Board, 1937, 152–158.

Petty, K., Kwon, J., & Skabardonis, A., 2006. *APeMS: An Arterial Performance Measurement System*. Presented at 85th Annual Meeting of the Transportation Research Board, Washington, D.C.

TRIP, 2015. *Washington's Top Transportation Challenges: Meeting the State's Need for Safe, Efficient Mobility and Economic Vitality*. http://www.tripnet.org/docs/WA_Transportation_Challenges_TRIP_Report_April_2015.pdf

Tufte, K.A., Bertini, R.L., Chee, J., Fernandez-Moctezuma, R.J., Periasamy, S., Matthews, S., Freeman, N., Ahn, S., 2010. *Portal 2.0: Toward a next-generation archived data user service.* Presented at 89th Annual Meeting of the Transportation Research Board, Washington, D.C.

VanDaniker, M.R., & Pack, M.L., 2009. *Visualizing real-time and archived traffic incident data*. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

Wang, Y.H., Corey, J., Lao, Y.T., & Wu, Y.J., 2009. *Development of a statewide online system for traffic data quality control and sharing*. Project Report 61-6022. Transportation Northwest (TransNow), Seattle, Washington.

Washington, S., Haque, M., Oh, J., Lee, D., 2014. *Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots*. Accident Analysis and Prevention, 66, 136-146.

Washington Traffic Safety Commission, 2016. *Washington State Strategic Highway Safety Plan 2016*. Washington State Department of Transportation.

Washington Traffic Safety Commission, 2013. *Washington State Strategic Highway Safety Plan 2013*. Washington State Department of Transportation

Wongsuphasawat, K., Pack, M.L., Filippova, D., VanDaniker, M.R., & Olea, A., 2009. *Visual analytics for transportation incident data sets*. Journal of the Transportation Research Board, 2138, 135-145.

Wu, Y.J., An, S., Ma, X.L., & Wang, Y.H., 2011. *Development of a web-based analysis system for real-time decision support on arterial networks*. Journal of the Transportation Research Board, 2215, 24-36.

Wu, Y.J., & Wang Y.H., 2009. *An interactive web-based system for urban traffic data analysis*. International Journal of Web Applications, 1(4), 241-252.

Wu, Y.J., Wang, Y.H., & Qian, D., 2007. *A Google-map-based arterial traffic information system*. Proc., ITSC 2007: IEEE Intelligent Transportation Systems Conference, Seattle, Wash., IEEE, Piscataway, N.J., 968–973.

Xie, G., & Hoeft, B., 2012. *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, pp. 45-56.

Zeng, Z.Q., Zhu, W.B., Ke, R.M., Ash J., Wang, Y.H., Xu J.P., Xu, X.X., 2017. *A generalized nonlinear model-based mixed multinomial logit approach for crash data analysis*. Accident Analysis and Prevention, 99, 51-65.