12-1-2016

A Business Intelligence Framework for Network-level Traffic Safety Analyses

Naveen Kumar Veeramisti
University of Nevada, Las Vegas, n.veeramisti@gmail.com

A BUSINESS INTELLIGENCE FRAMEWORK FOR NETWORK-LEVEL TRAFFIC SAFETY ANALYSES

By

Naveen Kumar Veeramisti

Bachelors of Civil Engineering
Crescent College of Engineering, Chennai
2001

Masters of Science in Engineering
University of Nevada, Las Vegas
2007

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

Department of Civil and Environmental Engineering and Construction
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
December 2016

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

November 18, 2016

This dissertation prepared by

Naveen Kumar Veeramisti

entitled

A Business Intelligence Framework for Network-level Traffic Safety Analyses

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Engineering – Civil and Environmental Engineering
Department of Civil and Environmental Engineering and Construction

Alexander Paz, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

Mohamed Kaseko, Ph.D.
Examination Committee Member

Moses Karakouzian, Ph.D.
Examination Committee Member

Hualiang Teng, Ph.D.
Examination Committee Member

Brendan Morris, Ph.D.
Graduate College Faculty Representative

ABSTRACT

A Business Intelligence Framework for Network-level Traffic Safety Analyses

by

Naveen Kumar Veeramisti

Dr. Alexander Paz, Examination Committee Chair
Associate Professor, Civil and Environmental Engineering
University of Nevada, Las Vegas

Currently, there are both methodological and practical barriers that together preclude a substantial use of theoretically sound approaches, such as the ones recommended by the Highway Safety Manual (HSM), for traffic safety management. Although the state-of-the-art provides theoretically sound approaches such as the Empirical Bayes method, there are still various important capabilities missing. Methodological barriers include among others (i) lack of a theoretically sound approach for corridor-level network screening, (ii) lack of a comprehensive approach for estimation of Safety Performance Functions based on a simultaneous consideration of both crash patterns and associated explanatory variables, and (iii) lack of theoretically sound methods to forecast crash patterns at the regional level. In addition, the use of existing theoretically sound approaches such as the ones recommended by the HSM are associated with important practical barriers including 1) significant data integration requirements, 2) a special schema is needed to enable analysis using specialized software, 3) time-consuming and intensive processes are involved, 4) substantial technical knowledge is needed, 5) visualization capabilities are limited, and 6) coordination across various data owners is required.

Considering the above barriers, most practitioners use theoretically unsound methodologies to perform traffic safety analyses for highway safety improvement programs. This

research proposes a single comprehensive framework to address all the above barriers to enable the use of theoretically sound methodologies for network wide traffic safety analyses. The proposed framework provides access through a single platform, Business Intelligence (BI), to theoretically sound methods and associated algorithms, data management and integration tools, and visualization capabilities. That is, the proposed BI framework provides methods and mechanisms to integrate and process data, generate advanced and theoretically sound analytics, and visualize results through intuitive and interactive web-based dashboards and maps.

The proposed BI framework integrates data using Extract-Load-Transform process and creates a traffic safety data warehouse. Algorithms are implemented to use the data warehouse for network screening analysis of roadway segments, intersections, ramps, and corridors. The methodology proposed and implemented here for corridor-level network screening represents an important expansion to the existing methods recommended by the *HSM*. Corridor-level network screening is important for decision makers because it enables to rank corridors rather than sites so as to provide homogenous infrastructure to minimize changes within relatively short distances. Improvements are recommended for long sections of roadways that could include multiple sites with the potential for safety improvements. Existing corridor screening methodologies use observed crash frequency as a performance measure which does not consider regression-to-the-mean bias. The proposed methodology uses expected crash frequency as a performance measure and searches corridors using a sliding window mechanism which addresses crash location reporting errors by considering the same section of roadway multiple times using overlapping windows.

The proposed BI framework includes a comprehensive methodology for the estimation of SPFs considering simultaneously local crash patterns and site characteristics. The current state-

of-the-art uses predefined crash site types to create single clusters of data to generate regression models, SPFs, for the estimation of predicted crash frequency. It is highly unlikely for all crash sites within a single predefined cluster/type to have similar crash patterns and associated explanatory characteristics. That is, there could be sites within a cluster/type with different crash patterns and explanatory characteristics. Hence, assigning a single predefined SPF to all sites within a type is not necessarily the best approach to minimize the estimation error. To address this issue, a mathematical program was formulated to determine simultaneously cluster memberships for crash sites and the corresponding SPFs. Cluster memberships are determined using both crash patterns and associated explanatory variables. A solution algorithm coupling simulation annealing and maximum log likely estimation was implemented and tested. Results indicated that multiple SPFs for a crash and/or facility type can maximize the probability of observing the available data to increase accuracy and reliability. The estimated SPFs using the proposed approach were implemented within the BI framework for network screening. The results illustrate that the gain in predicted crashes provided by the SPFs translates into superior rankings for sites and corridors with the potential for safety improvements.

A performance-based safety program requires the forecasting, at the regional level, of safety performance measures and establish targets to reduce fatalities and serious injuries. This is in contrast to the analysis required for traffic safety management where forecasts are required at the site or corridor level. For regional level forecasting, historically, theoretically unsound methods such as extrapolation or simple moving-average models have been used. To address this issue, this study proposed deterministic and stochastic time series models to forecast performance measures for performance-based safety programs. Results indicated that stochastic

time series, a seasonal autoregressive integrated moving average model, provides the required statistically sound forecasts.

In summary, the fundamental contributions of this research include: (i) a theoretically sound methodology for corridor level network screening, (ii) a comprehensive methodology for the estimation of local SPFs considering simultaneously crash patterns and associated explanatory variables, and (iii) a theoretically sound methodology to forecast performance measures to set realistic targets for performance-based safety programs. In addition, this study implemented and tested the above contributions along with existing algorithms for traffic safety network screening within a single BI platform. The result is a single web-based BI framework to enable integration and management of source data, generation of theoretically sound analyses, and visualization capabilities through intuitive dashboards, drilldown menus, and interactive maps.

ACKNOWLEDGEMENTS

I would first like to thank Dr. Alexander Paz for everything that he has done for me both as my academic advisor and life mentor. His knowledge, care and impeccable insights have helped me grow during my academic training. He acted as my role model and provided me with invaluable guidance and inputs throughout my doctoral studies at University of Nevada, Las Vegas. I cherished the time spent with him on learning about academic life. I am also grateful to my other committee members, Dr. Mohamed Kaseko, Dr. Hualiang Teng, Dr. Moses Karakouzian and Dr. Brendan Morris for guiding my research in the right direction with their expertise and insightful comments during proposal defense and one-on-one meetings. Thank you very much for your time and support. Many thanks to Mrs. Julie Longo for her help, as a technical writer, reviewing my research papers which are part of this dissertation.

The research was supported by University of Nevada, Las Vegas and Nevada Department of Transportation (NDOT) which also provided data and useful information. I would like to express my sincere gratitude to my friend and colleague, Justin Baker, for his valuable time and support during my research. I would sincerely thank my colleagues in the Transportation Research Center, Romesh Khaddar, Pankaj Maheshwari, Indira Khanal, Mukesh Khadka, Cristian Artega, Daniel Emaasit, and John Bertini for their valuable time, support and friendship. During this challenging journey, they were the sources of encouragement and motivation. I also would like to thank interns, Maxime Vey, Kenny Moupita, Simon Geslot, Alexandre Chevaucher, Romain Hamon, and Matthew Raybuck for their assistance during my doctoral studies.

Most importantly, I would like to express my deepest gratitude to my family, especially to my parents, brother, grandmother, aunt and uncle, sisters and parents-in-laws for their

enduring love and support. And very special thanks to my wife, Krithika, for fulfilling my life as my soul mate and my lifelong best friend, and always showing me unconditional love and providing immense support in achieving career goals. She provided encouragement and motivation when I needed it most. I should thank my stress buster, companion and beloved son, Vidyuth for his understanding during my unavailability to play with him. I greatly appreciate my extended family and friends for the support they provided throughout this journey and for understanding my devotion to this endeavor. Special thanks to Karthik Collinjivadi and Kapil Kulakunnath, who are my friends, role models, and mentors. I would like to thank them for being such a positive influence in my life and on my educational path.

DEDICATION

To my spouse Krithika and my son Vidyuth for their unconditional love, support, and encouragement throughout my doctoral studies.

Especially to my Mom Mrs. V.K. Lakshmi Rajam for her dream and hard work which made my studies continue to this level.

Also to my dad Mr. V.S. Kotteeswarudu for his support, brother Niran for his companionship and friends for their advice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Significant resources invested on roadway safety management have not translated yet into less traffic crashes in the United States (NHTSA, 2013). Such legislation as the Safe Accountable Flexible Efficient Transportation Equity Act – A Legacy for Users (SAFETEA-LU) and the Moving Ahead for Progress in the 21st Century (MAP-21) mandate states to develop comprehensive Highway Safety Improvement Plans (HSIPs) for improving transportation safety (FHWA, 2013; SAFETEA-LU, 2005).

Two critical programs that are part of HSIP are: 1) an annual report of highway locations with the most severe traffic safety needs and 2) an annual report of a performance-based safety program (HSIP, 2015). The purpose of the first program is to identify the most hazardous site locations that can be improved effectively by implementing countermeasures. This process – termed as a roadway safety management process by Part B of the *Highway Safety Manual* (*HSM)* – includes four steps: network screening, diagnosis and countermeasure selection, economic appraisal and priority ranking, and countermeasure evaluation (AASHTO, 2010). Network screening involves the systematic identification and raking of site locations with potential for safety improvements. The top ranked sites are investigated through the other three steps listed above (AASHTO, 2010; AASHTO, 2011).

The purpose of the second program is to reduce statewide fatalities and serious injuries by setting a target every year and achieving it for four safety performance measures: 1) number

of fatalities, 2) number of serious injuries, 3) rate of fatalities, and 4) rate of serious injuries (FHWA, 2015).

The existing literature reveals that practitioners continue using theoretically unsound methods to conduct traffic safety analysis to develop HSIPs (FHWA, 2015). Methods such as crash rates, crash frequency, equivalent property damage only, crash severity index have limitations including the bias associated with the volume, segment length, and regression-to-the-mean as well as incorrect model forms and lack of reliability measures (AASHTO, 2010). For annual reporting of highway locations exhibiting the most severe traffic safety needs, in FY 2014, only four states in the U.S. used theoretically sound methodologies for network screening such as the Empirical-Bayes (EB) methodology provided in the *HSM* (FHWA, 2015). The EB methodology, involves the use of safety performance functions (SPFs) to estimate the number of crashes that would be expected in the analysis period at locations with traffic volumes and other characteristics similar to the one being analyzed (Montella, 2010). The predicted crash estimates obtained using SPFs are then combined, using weights, with the observed count of crashes to obtain a better estimate of the expected number of crashes. The weighted adjustment accounts for the reliability of the safety performance function that is applied. Crash estimates produced using safety performance functions with over-dispersion parameters that are low (which indicates higher reliability) have a larger weighted adjustment. Larger weighting factors place a heavier reliance on the SPF to predict the long-term predicted average crash frequency per site (AASHTO, 2010).

For the performance-based safety program, states use such aspirational targets as promoting zero fatalities/deaths, the American Association of State Highway and Transportation Officials (AASHTO) target to halve fatalities by 2030, or aim for a gradual (e.g., 3%) decrease in

fatalities every year (NCHRP, 2010). In addition, these targets are instituted based on discussions of focus groups in state agencies. Few states use trends to forecast performance measures and set targets.

Both methodological and practical barriers together preclude a substantial use of theoretically sound approaches for traffic safety analysis and management. Although the state-of-the-art provides theoretically sound approaches, there are still various important capabilities missing. Methodological barriers include among others (i) lack of a theoretically sound approach for corridor-level network screening, (ii) lack of a comprehensive approach for estimation of Safety Performance Functions based on a simultaneous consideration of both crash patterns and associated explanatory variables, and (iii) lack of theoretically sound methods to forecast crash patterns at the regional level. In addition, the use of existing theoretically sound approaches are associated with important practical barriers including 1) significant data integration requirements, 2) a special schema is needed to enable analysis using specialized software, 3) time-consuming and intensive processes are involved, 4) substantial technical knowledge is needed, 5) visualization capabilities are limited, and 6) coordination across various data owners is required (Alluri, & Ogle, 2012; Tarko et al., 2014; Paz et al., 2015c).

As a consequence of these barriers, there is a significant gap between state-of-the-art and the state-of-the-practice methodologies at federal, state, and local level. The ability to adopt and use theoretically sound methodologies by practitioners for traffic safety analysis is key for improving traffic safety. Tools and methods from informatics and computer science, such as data warehousing and business intelligence (BI), provide opportunities to develop a data warehouse connected to the source data as well as to the required analytical models (Chen et al., 2012; Rittman, 2013).

In order to facilitate the development of annual reporting of highway locations exhibiting the most severe traffic safety needs, this research seeks to address the barriers listed above to enable practitioners to use theoretically sound methodologies for traffic safety management. Mechanisms were investigated and developed to deploy, enhance, and facilitate the use of theoretically sound algorithms for the development of HSIPs.

1.2 Objectives of the Dissertation

The primary objectives of this dissertation include the development of: (i) a theoretically sound approach for corridor-level network screening, (ii) a comprehensive approach for estimation of SPFs based on a simultaneous consideration of both crash patterns and associated explanatory variables, and (iii) a theoretically sound method to forecast crash patterns at the regional level. In addition, existing algorithms for network screening from the HSM and the above developments are together implemented within a single BI framework to enable: 1) access, integration and management of data, 2) analyses, and 3) visualization of results.

Addressing the above objectives required reimplementation and expansion of the network-screening process discussed in Part B of the *HSM* (AASHTO, 2010) to provide a single framework for data processing, integration, analysis and visualization. Algorithms were coded for network screening analysis of roadway segments, intersections, ramps, and corridors. The algorithm proposed and implemented here for corridor level network screening represents an important expansion to the existing methods recommended by the HSM. In practice, screening and analysis/ranking at the corridor level for the entire network is required for various reasons including the need to provide as homogeneous as possible conditions across the roadway network. Homogeneous conditions are associated with less driving distractions, surprises, or confusion.

As described above, appropriate SPFs are essential to determine reliable estimates of predicted crash frequency for network screening (AASHTO, 2010; Hauer, 2015). SPFs are crash prediction models represented by mathematical equations that relate the number of crashes to site characteristics. In the context of traffic safety, typically, analysis sites are grouped into site subtypes based on predefined characteristics. SPFs for crash severity within these subtypes are available (AASHTO, 2010; Lu et al., 2013). However, in reality, it is unlikely for all the sites in a single site subtype (single cluster) to have a similar crash pattern as a function of pre-defined explanatory characteristics. Hence, it is possible to have different clusters in terms of crash patterns within pre-defined site subtypes. To address this issue and estimate superior local SPFs, a clusterwise regression approach (Lau et al., 1999) was developed and implemented as part of the proposed BI framework. A mathematical program was formulated to estimate simultaneously parameters of SPFs and assign sites to appropriate cluster (SPF) based on crash patterns and associated explanatory characteristics.

To facilitate annual reporting of a performance-based safety program, a time- series methodology was developed and implemented within the proposed framework to forecast statewide traffic safety performance measures and set targets. With actual crash data from a source database and a sound statistical approach, forecasts using time-series models are likely to lead to reasonable targets for the performance measures. These targets can be used to determine future statewide safety improvement programs and policies. From the perspective of state agencies, predicting the number of fatalities and serious injuries is significantly important to meet the requirements of MAP-21 (FHWA, 2013).

Business Intelligence provides methods and mechanisms to process data and generate advanced analytics as well as interactive and intuitive visuals. The proposed BI framework

integrates data using Extract-Load-Transform process and creates a traffic safety data warehouse. A Business Intelligence approach is adopted to automate data integration and management and to connect analysis to the source data. The proposed approach for implementation involved an Oracle Data Integrator (ODI) (Dupupet et al., 2013), Oracle R Enterprise (McDermid & Taft, 2014), and Oracle Business Intelligence Enterprise Edition (OBIEE) (Rittman, 2013). Each year when new data gets loaded into the source database, the data warehouse is updated automatically by means of an extract-load-transform (ELT) process for further analysis. Oracle R accesses the data for analytical modeling and OBIEE provides results with intuitive visual graphics and maps using BI dashboards. By using this framework, the analyst can perform customized analyses. Dashboards can be used to trigger special-purpose analyses and tasks based on input parameters. With minimal effort, practitioners can analyze and view results using an interactive web-based interface powered with drill downs. Figure 1.1 provides a conceptual illustration of the proposed framework.

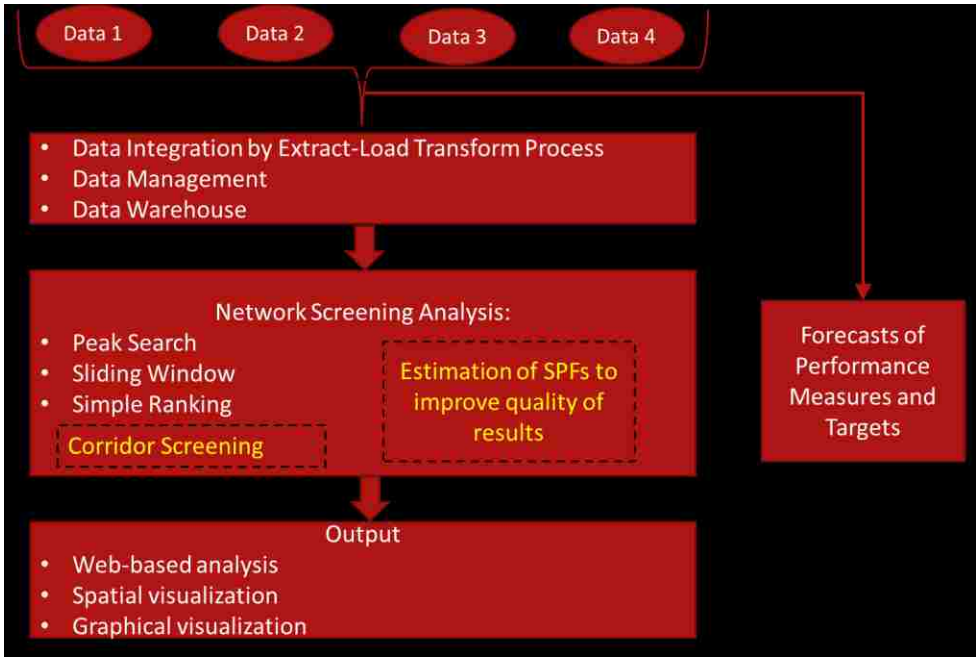


Figure 1.1 A Framework for Traffic Safety Analysis.

Specific aspects required to address the objectives of this research include:

1) The development of a comprehensive database system that enables the use and development of theoretically sound methodologies for traffic safety analysis. The database is developed to take advantage of existing tools such as Safety Analyst, a state-of-the-art software for Traffic Safety Management. The main tasks required to develop the database system include: i) identify the data sources, ii) develop the systems and tools to integrate them, iii) develop the databases consistent with the Safety Analyst format, iv) use Safety Analyst for analysis, and v) develop systems for visualizing the results. The results from Network Screening analysis using Safety Analyst are used for quality control of the proposed concept in Figure 1.1. Safety Analyst represents a traditional approach which involves multiple individual steps and such components as customized data management and visualization tools a, making the process complicated and time consuming. Figure 1.2 illustrates this traditional approach for traffic safety analysis.

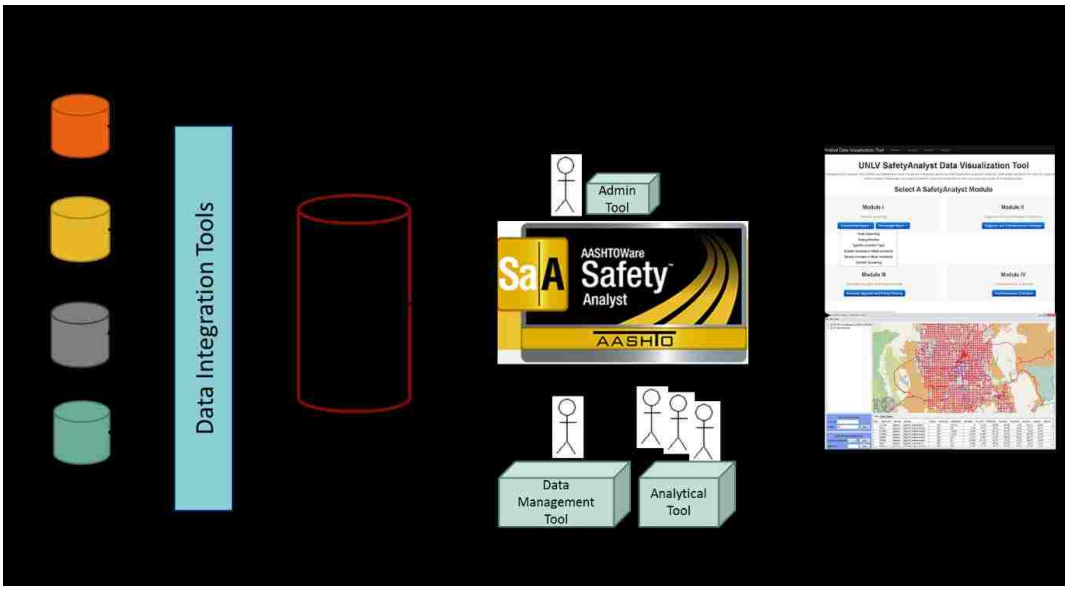


Figure 1.2 Traditional Approach for Traffic Safety Analysis.

2) Reimplement network screening algorithms from the *Highway Safety Manual* within the proposed BI framework to provide a single platform for data integration, management, analysis, and visualization. For illustration purposes, this reimplementation is performed using Oracle R Enterprise, within Oracle Business Intelligence Enterprise Edition. With this framework, as illustrated in Figure 1.3, network screening can be performed on a web-based interface with a data warehouse directly connected to the source. This proposed approach represents a paradigm shift where theoretically sounds methodologies are available to practitioners through a platform that addresses all existing barriers which have prevented their use.

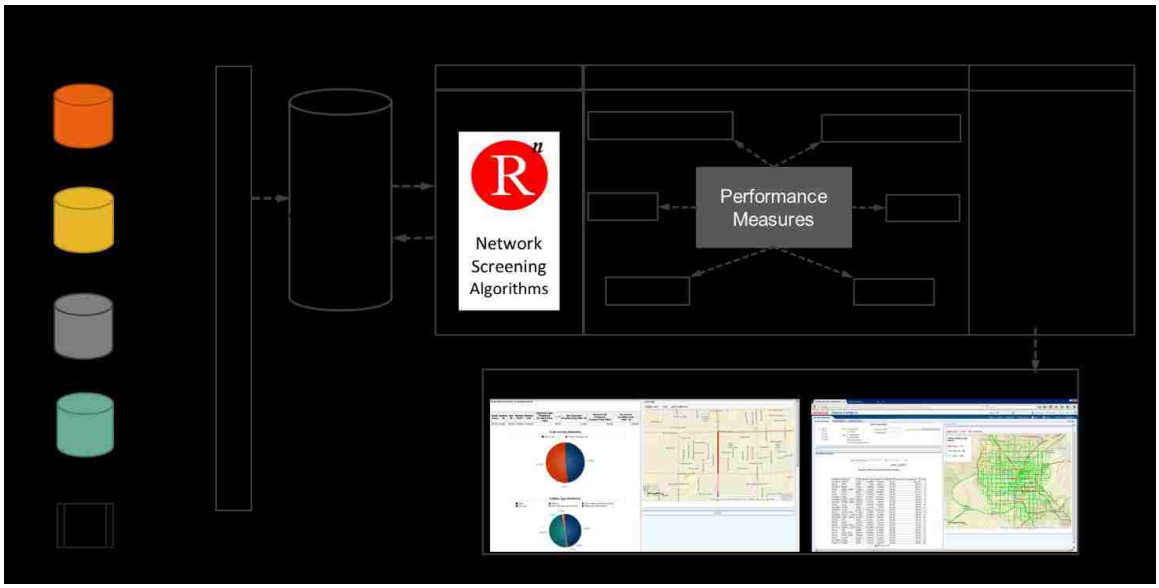


Figure 1.3 Business Intelligence Approach for Traffic Safety Analysis.

3) Develop and implement, within the proposed BI framework, a methodology for corridor level network screening for potential safety improvements using an expected crash frequency as the performance measure. This is important for incorporating safety measures into corridor planning studies. Corridor-level network screening provides the capability

to compare the safety performance of extended corridors rather than comparing the safety performance of individual sites. A corridor may consist of multiple roadway segments, intersections, and/or ramps, which are aggregated together and analyzed as a single entity. In this study, two types of corridor-level network screening methods are considered: 1) Fixed Corridor screening, and 2) Corridor Search.

4) Formulate a mathematical program in the context of clusterwise regression to estimate simultaneously parameters of SPFs and assign crash sites to appropriate clusters used for this estimation. Develop a solution algorithm to obtain appropriate cluster memberships of sites associated with SPF and its parameters. Multiple SPFs would maximize the probability of observing the available data to increase accuracy and reliability. In existing safety literature, the proposed clusterwise regression approach is first of its kind to estimate SPFs. Crash sites in the entire data are clustered to estimate SPFs by maximizing the log-likelihood of a Negative Multinomial distribution function. The decision variables are the number of SPFs, the parameters of Negative Multinomial SPFs, and the cluster memberships.
5) Develop and implement a methodology for forecasting traffic safety performance measures as required by MAP-21. This methodology could be used by transportation agencies to set and achieve realistic targets for performance- based traffic safety programs. By using the methodology, agencies can forecast and report targets easily every year. Deterministic and stochastic time series models were tested using four independent and univariate time series from the crash data collected by the Nevada Department of Transportation. Baseline forecasts from the time series models can further be used to estimate the reduction of fatalities and serious injuries to set realistic targets.

1.3 Organization of the Dissertation

This dissertation is divided into six chapters and follows a manuscript format with this chapter as introduction. In Chapter 2, "Development of a Comprehensive Database System for Safety Analyst", a comprehensive database system and tools to provide data to multiple traffic safety applications, was developed (Paz et al., 2015c). A number of data management tools were developed to extract, collect, transform, integrate, and load the data. The system includes consistency-checking capabilities to ensure the adequate insertion and update of data into the database. This system focused on data from roadways, ramps, intersections, and traffic characteristics for Safety Analyst. To test the proposed system and tools, data from Clark County, which is the largest county in Nevada and includes the cities of Las Vegas, Henderson, Boulder City, and North Las Vegas, was used. The database and Safety Analyst together helped to identify the sites with the potential for safety improvements.

Chapter 3, "A Business Intelligence Framework for Traffic Safety Network Screening", includes a new methodology for corridor level network screening as well as the implementation of existing algorithms to provide a single site analysis framework. The results obtained using proposed methodology were compared with the state-of-the-art for corridor screening. Similarly, results from the reimplemented algorithms were compared with those obtained using Safety Analyst (AASHTO, 2010).

Chapter 4, "Estimation of safety performance functions using clusterwise regression", proposes a mathematical program to assign c r a s h sites to clusters and simultaneously seek sets of parameter values for the corresponding SPFs so as to maximize the probability of observing the available data. A simulated annealing coupled with maximum likelihood estimation was used to solve the mathematical program. Results were analyzed for two site

subtypes 1) roadway segments for urban multi-lane divided arterials, and 2) urban 4-leg signalized intersections. The proposed approach improved the predicted number of crashes with multiple SPFs within the same site subtype. In addition, network screening results using the proposed SPFs illustrate substantial differences compared to those obtained from predefined cluster of crash sites.

In Chapter 5, "Forecasting Traffic Safety Performance Measures with Deterministic and Stochastic Time Series Models", traffic safety performance measures were forecasted to facilitate the reduction of fatalities and serious injuries. Given the lack of exposure data (e.g., traffic counts), time series were used to conduct the forecasting. Deterministic and stochastic models were developed using four independent and univariate time series from the crash data collected by the Nevada Department of Transportation. Results indicated that the seasonal autoregressive integrated moving average (SARIMA) model provided the best forecast measures for the data.

Chapter 6 summarizes the conclusions gained from this research, identifies significant contributions, and recommends potential future research directions.

CHAPTER 2

DEVELOPMENT OF A COMPREHENSIVE DATABASE SYSTEM FOR SAFETY ANALYST

2.1 Introduction

The NHTSA's Highway Traffic Safety Grants for Fiscal Year (FY) 2013 were estimated to be $643 million (NHTSA, 2013). In spite of enormous resources spent on highway traffic safety, motor vehicle crashes are of critical concern in the United States. Based on statistical projections from NHTSA's Fatality Analysis Reporting System (FARS), traffic fatalities increased from 32,367 in 2011 to 34,080 in 2012, a 5.3% increase. In fact, 2012 was the first year since 2005 to have a year-to-year increase in fatalities, which indicates that considerable work is needed to improve highway safety (NHTSA, 2012).

FHWA's Highway Safety Improvement Program (HSIP) is a critical component of the safety provisions in Moving Ahead for Progress in the 21st Century Act (MAP–21, P.L. 112-141) (FHWA, 2013). As a part of HSIP, state Departments of Transportation (DOTs) developed a Strategic Highway Safety Plan (SHSP) to identify, analyze, and address traffic safety problems. State-of-the-art tools have been created to support the development of SHSP and generate better traffic solutions for existing and emerging safety problems. Some of these tools include the Interactive Highway Safety Design Model (IHSDM, 2010), the *Highway Safety Manual* (HSM), and the software tool, Safety Analyst. These tools can be used by DOTs to satisfy MAP-21's performance-based federal program, which mandates that state DOTs to establish safety

performance targets and achieve them within two years (FHWA, 2013). This requires a program for highway safety management that should include:

1) Identifying hazardous locations,
2) Diagnosing identified hazardous locations and countermeasure selections,
3) Estimating the cost of the countermeasures, and
4) Estimating the benefits of the countermeasures.

These new tools address many limitations of traditional safety analysis tools, including bias associated with volume, segment length, and regression-to-the-mean as well as incorrect model forms and lack of reliability measures (HSM, 2010; American Association of State Highway and Transportation Officials.; Montella, 2010; Alluri, 2010; Hauer, 1997; iTRANS, 2003). In order to address these limitations, state-of-the-art tools, including Safety Analyst, use analytical methods that require comprehensive datasets in order to provide sufficient information and to capture intricate spatio-temporal characteristics and interactions in the traffic system.

In their FY 2013 budget estimate, NHTSA determined that data-driven, self-sustaining highway safety programs needed to be developed and implemented to reduce highway injuries and fatalities (NHTSA, 2013). The federal government has spent considerable resources to build accurate and timely safety datasets at the national and state levels (Alluri, 2008). Key safety data include information about crashes, roadways, traffic flow, driver history, citation/adjudication, and construction projects (HSM, 2010; Ogle, 2007). This data is required by a number of other safety programs, such as the Highway Rail Grade Crossing Program and the High Risk Rural Road Program. Currently, various divisions at many state DOTs collect and maintain datasets

based on their data needs; some of this data is shared across divisions. However, this approach may not be the best for a number of following reasons:

a. Not all interested groups are aware of the availability of data at each division.
b. There is no consistency in terms of how the information is stored and the data normalized.
c. Typically, the datasets are developed without explicitly considering the needs of the various applications used by different divisions.
d. New emerging tools, such as Safety Analyst, require data to be collected from multiple divisions; additionally, these tools need data that typically is not available.
e. The training of traffic safety engineers and professionals on the use of new applications, such as Safety Analyst, requires the corresponding applications to be ready for use with all the necessary data available.
f. Coordination with other statewide public safety agencies requires a comprehensive approach to integrate and enable access to the data as well as to provide maintenance capabilities.

A comprehensive approach using state-of-the-art tools is required to collect data and manage existing data needs, which are significant, as well as to develop better solutions. The literature indicates data collection and integration methods for transportation applications have been developed, including frameworks for geographic information systems (GIS) (Ziliaskopoulos & Waller, 2000; Khan et al, 2010; DIP, 2001; Dueker & Butler, 1998; Devogele, 1998; Vonderohe, 1998; Pendyala, 2008), database/data-warehouse systems (O'Packi et al., 2000; Ming & Lei 2010; Pack et al., 2008; Hall et al., 2005), and visualization tools (Gan et al., 2012; Qin et al., 2011; Wu, Wang & Qian, 2007). However, most DOTs do not have access to a

comprehensive database system that enables them to take full advantage of existing tools, including Safety Analyst. With such a database system, agencies would be able to develop safety performance functions (SPFs) that are jurisdiction-specific to better estimate performance measures. Previous studies show that methodologies to develop such systems were relatively limited.

Many state DOTs are rich in data. However, it is a herculean task to identify the data sources, develop the systems to integrate them, and develop the databases. This study developed a database and visualization system for traffic safety engineering, designed to provide data to multiple transportation applications. Recently, the development focused on providing data and visualization capabilities for Safety Analyst. However, a recent nationwide survey revealed major deterrents in using Safety Analyst (Xiao Qin, 2011), including the unavailability of comprehensive data sources and tedious methods for data importing and processing. This study developed a database system tailored to Safety Analyst specifically for traffic safety analysis in Clark County, Nevada. However, all the tools developed to create the database system could be used to create similar databases for other locations and/or to expand existing databases.

Figure 2.1 illustrates the conceptual framework for this database and visualization system. Raw data was processed using data management tools to create a comprehensive, normalized, and optimized database. View tools were used to provide the data required by each application, in the corresponding format and level of resolution. Visualization tools were used to provide multiple graphical representations of the inputs and outputs for each application. Many analysis tools exist, including Safety Analyst that do not provide visualization capabilities. This was a significant limitation, considering the spatial nature of the problem.

2.2 Safety Analyst

Safety Analyst provides a suite of analytical tools to identify and manage system-wide safety improvements (AASHTO, 2011). Safety Analyst uses an empirical Bayes (EB) method as an alternative to traditional safety analysis methods, such as frequency, rate, critical rate, or crash index. The EB approach provides a mechanism to address issues associated with bias, incorrect model form, and lack of a reliability measure that cannot be addressed using traditional methods (HSM, 2010; AASHTO, 2010; Montella, 2010; Alluri, 2010; Hauer, 1997; iTrans, 2003; Alluri, 2008). Safety Analyst consists of four tools: administration, data management, analysis, and countermeasure implementation. The administration tool includes federal, agency, and system components (Wu, Wang, Qian, 2007). The federal component provides access to the default site subtype definitions, countermeasure management, and national default SPFs. The agency component provides access to various operations, including adding, changing, and removing data attributes, with the exception of mandatory data attributes. Further, this component enables the modification of national SPFs with agency-specific SPFs. The system component maintains local or remote databases, and combines the database with the federal and agency components.

Local or remote databases can be imported using the data management tool (Wu, Wang, Qian, 2007). Currently, Safety Analyst supports two basic mechanisms for data import, a file import and database-to-database mapping. For DOTs that maintain a complete data inventory in a database management system (DBMS) compliant with structured query language (SQL), the database-to-database mapping mechanism is the best alternative to load data into Safety Analyst. The view tools developed in this study provides this feature. For DOTs that do not maintain a database with all the required data for Safety Analyst, the data management tools developed in this study can be used to generate a DBMS with all the required data.

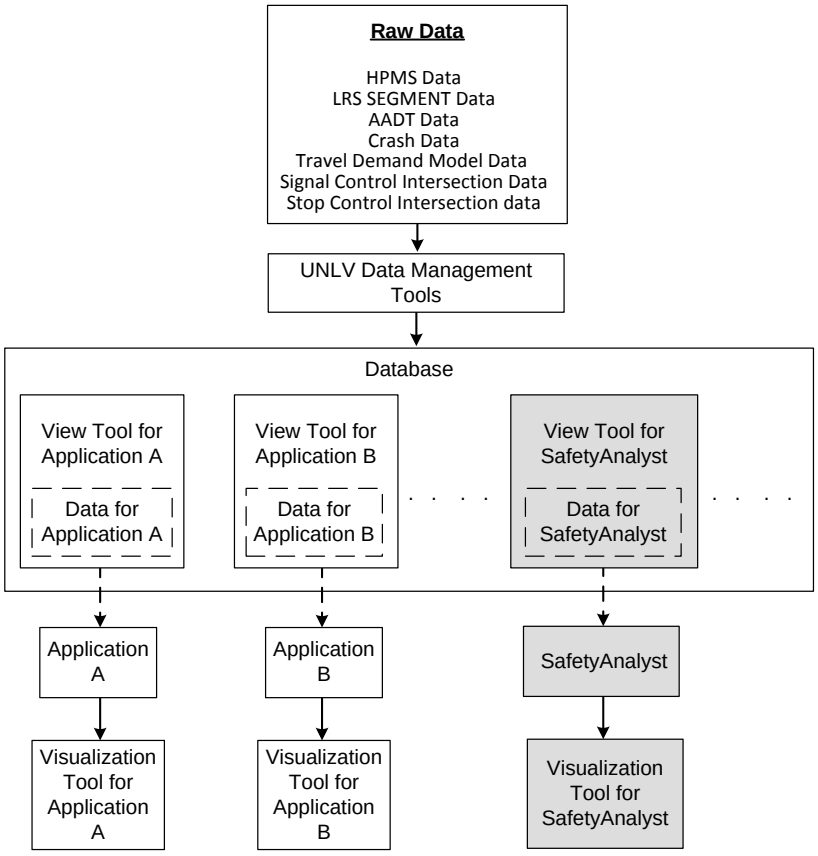


Figure 2.1 Conceptual Framework for the Comprehensive Database and Visualization System Developed in this Study.

The file import is a less desirable mechanism because it does not provide all the capabilities of having the data in a DBMS. Safety Analyst supports data inventory files in extensive mark-up language (xml) and comma-separated value (csv) formats. However, the inventory files have to satisfy a particular format. It is unlikely that DOTs have readily available xml or csv datasets that satisfy the required format. Hence, developing a DBMS for Safety Analyst is recommended.

The analysis tool, used to perform various analyses (Wu et al., 2007), has a set of four modules, including (HSM, 2010; Wu et al., 2007):

a. *Network Screening Module:* identifies and ranks sites using the EB method for potential safety improvements.

b. *Diagnosis and Countermeasure Selection Module:* helps to diagnose safety problems at specific sites using answers provided by the user for a set of built-in questions. Based on the diagnosis, the user can select countermeasures to reduce crash frequency and severity at specific sites.

c. *Economic Appraisal and Priority Ranking Module:* provides economic evaluation of a specific countermeasure for a specific site or several alternative countermeasures for multiple sites. Further, it provides priority ranking of sites and proposed improvement projects based on benefit and cost estimates.

d. *Implemented Countermeasure Module:* provides before/after evaluation of implemented safety improvements. Data for construction projects and implemented countermeasures are required. This data can be imported using the countermeasure implementation tool.

2.2.1 Data

Critical data to perform traffic safety studies include crash, roadway, control, and traffic flow. A comprehensive plan for data collection was developed to obtain available data from various state agencies in Nevada, based on the Model Minimum Uniform Crash Criteria (MMUCC) and the Model Inventory of Roadway Elements (MIRE) (Alluri, & Ogle, 2011; Paz et al., 2015c). Based on these guidelines, approximately 150 data attributes were necessary for the development of a comprehensive safety database. Not all the data was required by Safety Analyst, however in this study, a data dictionary was developed to explicitly identify the mandatory data for Safety Analyst (HSM, 2010; Wu et al., 2007), as shown in Figure 2.2.

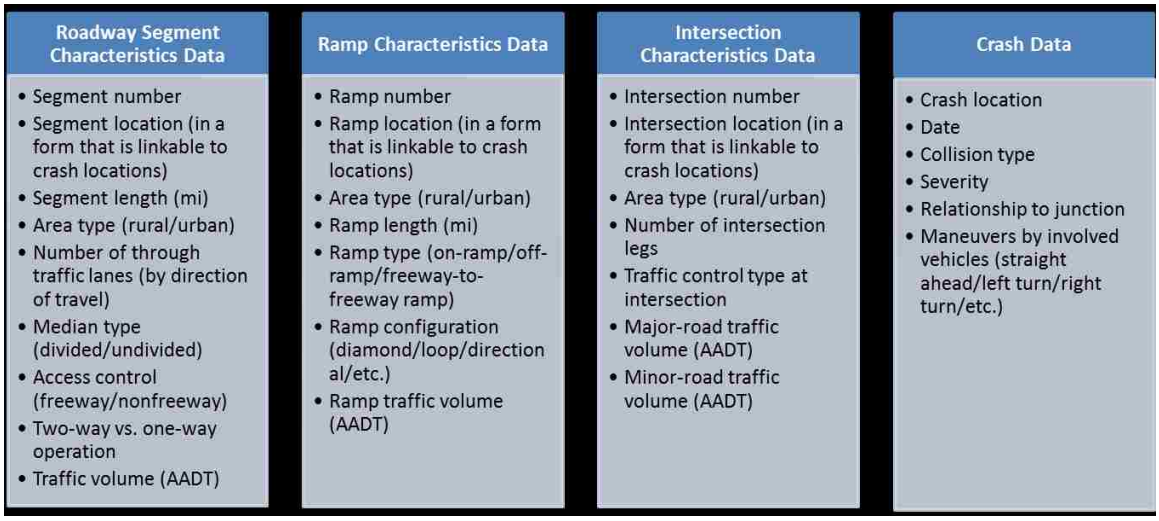


Figure 2.2 Mandatory Data Elements Required by Safety Analyst.

Most of the data in Figure 2.2 is available from various DOT sources, including FHWA's Highway Performance Monitoring System (HPMS); linear referencing systems (LRS) of road networks; travel demand models (TDM); and intersection, traffic volume, and crash datasets (Xiao Qin, 2011). For this study, data for roadway segments and ramps were obtained from the LRS, HPMS and TDMs. Crash data was obtained from the Nevada Accident and Citation Tracking System (NCATS). Annual average daily traffic (AADT) was collected from NDOT's Traffic Records Information Access (TRINA).

2.2.2 Road Network

A road network is the centerline map of routes in a GIS LRS. Most of the state DOTs have two levels of road networks, a state-level dataset (SDS) and a county-level dataset (CDS). The SDS can be used for federal aid and national highway system roads in Safety Analyst, and the CDS can be used for county-level minor arterial roads as well as for major and minor collector roads. Typically, an SDS road network is similar to an HPMS routes layer. When both SDS and CDS

road networks are unavailable, the HPMS routes layer in LRS (HPMS, 2010) can be used with some modifications.

For this study, the CDS road network in LRS was used, which included an additional system, Route Master identification (RMID), which is a unique identifier for referencing the route in the road network. The RMID improves the ability to reference the other data sources to the road network. Road network data includes the segment ID, RMID, type of road, county, begin and end milepost of the segment, cardinal direction, and length of the segment. The cardinal direction reflects the direction in which the road begins and ends.

2.2.3 HPMS data

The HPMS is a FHWA-maintained national-level system that includes data on the extent, condition, performance, use, and operating characteristics of the state-owned and some non-state-owned highways (SAUM, 2011). The HPMS data model by FHWA, which is in a GIS framework, provides the spatial relationships among data elements. FHWA mandates the state DOTs to submit complete, timely, and accurate HPMS data every year (SAUM, 2011). Hence, this data – integrated with other data sources – can be available to state DOTs for database development required for Safety Analyst.

For this study, Nevada HPMS data layers were used, including access control, facility type, functional classification, speed limit, through lanes, AADT, and urban code.

2.2.4 Travel Demand Model

Usually, urban metropolitan planning organizations (MPOs) have a GIS-based TDM for transportation planning and transportation improvement programs. The data from this model – such as number of lanes, speed limit, access control, functional classification, area code, travel

direction, one-way or two-way, and ramp configuration – can be used when HPMS data is not available. If a distinct county-level road network is not available, a TDM road network can be used for data on road segments, ramp segments, lengths, and mileposts.

For this study, the TDM of the Regional Transportation Commission of Southern Nevada (RTC-SN) was used to obtain data not available in the HPMS layers.

2.2.5 Crash Data

Every year, NHTSA spends much of its budget on their highway safety grants for the Crash Data Collection Program (NHTSA, 2013). The collection of crash data from states must be based on MMUCC guidelines. The crash data required by Safety Analyst is based on MMUCC guidelines as well. This study used located crashes (crashes with coordinates) and crash characteristics, for the years 2007 to 2011, from NCATS.

2.2.6 AADT Data

Safety Analyst requires AADT for all the segments to be used in a network-level analysis. Frequently, however, they are not available for all roadway classes. Typically, DOTs collect data to estimate AADTs for high functional classes of roads, such as freeways and state roads. Collecting similar data for arterials and local roads is an extensive and expensive process. This study used a simulation-based dynamic traffic assignment model, DynusT (DTA Primer, 2010; DynusT, 2008), to estimate AADT for locations with missing AADT for the latest year. These AADTs were projected for five years, using temporal factors developed from long-term counts.

2.2.7 Intersection Data

Typically, county agencies or metropolitan planning organizations (MPOs) have data for signalized intersections, including the location and type of control information. However, data

for stop-controlled intersections is not common, and needs to be collected. In this study, a methodology and a tool was developed to collect stop-control data efficiently. Signalized intersection data was obtained from the Freeway and Arterial System of Transportation (FAST), a division of RTC-SN.

Table 2.1 shows the source files typically available in state DOTs and/or MPOs as well as data in those files that are required by Safety Analyst. With this information, agencies can start collecting these files to develop a Safety Analyst database. Agencies can choose either HPMS files having data for road networks and crashes for roads maintained by state DOTs or HPMS files with data for road networks, TDMs, crashes, and intersections for county-level roads.

Table 2.1 Source Files and Their Data Elements to Build a Safety Database

Road Network	HPMS Files	TDM Model	Crash Data	Intersection
Segment ID	Routes	Travel Direction	Accident ID	Intersection ID
Ramp ID	Functional Classification	Functional Classification	Crash Location	Intersection Location
Segment Length	Access Control	Operation Way (1 or 2 Way)	Cash Date	Type of Control
Begin Milepost	Speed Limit	Speed Limit	Collision Type	Number of Legs
End Milepost	Through Lanes	Number of Lanes	Severity	
Route ID	Lanes_Left	County	Relationship To Junction	
County	Lanes_Right	Area Code	Direction of Involved Vehicle	
Increasing Milepost Direction	AADT	Ramp Configuration (sometimes available)	Maneuvers by Involved Vehicle	
	Urban	Ramp Type (sometimes available)		
	County	Segment ID		
		Ramp ID		
		Segment Length		

The road networks – along with HPMS files, including AADT and crash data – form an integrated database covering all state-owned roadways and ramp segments, at least. However, this study required data on county-level roads as well. Therefore, data from the CDS road network was integrated with data from HPMS layers, TDM, intersections, AADT, and crashes. During integration, some of the issues found among these datasets are as follows.

a. A spatial shift/gap exists among GIS shape files of various datasets, such as HPMS, CDS road network, and TDM layers.

b. No common ID exists among the HPMS, CDS road network, and TDM layers.

c. Segmentation lengths differ in HPMS layers and the CDS road network.

d. There is no unique RMID among the datasets.

e. Some data are incorrectly represented, such as ramp configurations and the number of lanes.

Certain issues in datasets are common because there is no consistency in data formatting and storage among divisions or departments. Furthermore, the collected data may or may not have been stored in the same geographical format, such as cardinal measurements, coordinate systems, and geometry. ArcGIS ModelBuilder (ArcGIS Geoprocessing, 2013) was used to develop the automated tools to solve these issues as discussed in the following section.

2.3 Data Management Tools

2.3.1 Data Collection Tools

Even though multiple data sources exist that provide a vast amount of the required data for Safety Analyst, various data attributes were missing or incomplete, including ramp type, ramp

configuration, and the type of control at intersections. Most of the missing data were collected using Google Earth, and the missing information was observed and coded in Google Earth as well. A data collection tool was developed to extract data as well as create ArcGIS shape files with all the information. This capability facilitated the development and integration of the database.

Safety Analyst requires that all collected data be integrated using either 1) a route and milepost; 2) a route, county and milepost; 3) a route, section, and distance; or 4) a section and distance. This study used a route and milepost index to integrate all the data because some of the datasets had this information. Although various commercial methods and tools are available (Ziliaskopoulos et al, 2000; Khan et al., 2010; DIP, 2001; Dueker & Butler, 1998; Devogele et al., 1998; Vonderobe et al., 1998; Pendyala et al., 2008) to integrate the data, integration tools using ArcGIS ModelBuilder were developed in this study to gain total control of the process and provide greater automation.

2.3.2 ArcGIS ModelBuilder Tool

ModelBuilder (ArcGIS Geoprocessing, 2010) is an application existing inside ArcGIS by which models can be created, edited, and managed. A model is built with a sequence of processes and data chained together. Once built, a model can be saved as a tool and embedded in an ArcMap toolbar. The two primary uses of ModelBuilder are to execute a process sequence that was created and to create additional tools with new capabilities. These tools can be launched from the tool dialog box or from Python scripts. Using the ModelBuilder tool, the following operations can be performed:

a. Change parameter values, such as buffer radius or tolerance limits, and re-run models;

b. Add more processes, such as components for a buffer or intersect, as well as data;

c. Delete processes and intermediate data; and

d. Visualize and explore the results in ArcMap.

ModelBuilder tools were developed to overcome all the data issues encountered with HPMS, road networks, TDM, and AADT. The three primary tools used are:

a. A mapping tool that maps road network segments spatially to data elements in HPMS, when there is geometry shift and no common field between them.

b. A linear referencing tool that creates a milepost index for each crash with respect to roadway segment, ramp, or intersection mileposts.

c. A dynamic segmentation tool that breaks/joins the segments at required locations.

2.3.3 Interface for Data Attribute Mapping

An interface for data attribute mapping was developed to populate the database, using data from existing sources. The interface established mapping for every attribute, and data source from user-file data attributes to corresponding database attributes in the database tables. This interface enables using existing data files without any modifications. The interface uses a Microsoft Excel spreadsheet (.xlsx), which is a metadata file with four columns. The first and second columns include the database table name and the attributes name, respectively. These names are fixed, and do not need to be changed. The third and fourth columns include the user's (agency) file name and attribute name, respectively.

Table 2.2 illustrates the metadata file. Only the user file name and user file attribute name have to be filled out by the user. Different DOTs store their data in various files, and common

unique IDs relate those files and attributes. For example, Nevada has roadway attributes in different files, such as CDS_Network, Las Vegas Median, HPMS_Access, and HPMS_SpeedLimit. Once a metadata file is filled in, the data-attribute mapping interface is used to insert and store data from the user file into the corresponding database tables and attributes.

Table 2.2 Sample of the Metadata File for Data Mapping

Database Table Name (Fixed)	Database Attribute Name (Fixed)	User File Name (to be filled by User)	User File Attribute Name (to be filled by User)
RoadwaySegment	agencySegmentID	CDS_Network	ID1
RoadwaySegment	beginLocation	CDS_Network	Beg_Route_
RoadwaySegment	endLocation	CDS_Network	End_Route_
RoadwaySegment	routeName	CDS_Network	Route_MAST
RoadwaySegment	routeType	CDS_Network	Route
RoadwaySegment	county	CDS_Network	County_code
RoadwaySegment	length	CDS_Network	Datum_Seg2
RoadwaySegment	terrain	HPMS_terrain	Terrain
RoadwaySegment	roadwayClass	CDS_Network	F08_FType1
RoadwaySegment	medianType	LasVegas_Median	MedianType
RoadwaySegment	accessControl	HPMS_Access	Value_name
RoadwaySegment	medianWidth	LasVegas_Median	MedianWidth
RoadwaySegment	postedSpeed	HPMS_SpeedLimit	PostedSpeed

2.3.4 Data Installation and Insertion

Inputs to an existing database can be either new or an update of previously inserted data. A data instantiation and insertion tool was developed to input data into the database, taking into consideration the interdependencies of the data. Input files were streamed and parsed with a Simple application programming interface (API) for an XML parser, also known as SAX, is used to store the data in a matrix.

When a row is read in the matrix, a ‘select’ query is performed on the database to determine existence of the object. If the object exists, an update is performed: a java object is

instantiated and its fields are updated with the values in the input file. Then, the update method of this java object is processed to update the database. If there is no matching object, a new object is instantiated and inserted into an 'EntityManager' class. Once all files are parsed and all the objects instantiated, the data can be inserted into the database. The 'EntityManager' class handles the priorities of the tables automatically in order to satisfy interdependencies between tables. Inputting data without using 'EntityManager' might lead to either data insertion failure or database corruption due to violation of table interdependencies. For example, accident vehicle data is dependent on accident data, and a mechanism is required to account for this dependency.

2.4 Database Schema

The database schema provides the structure of a DBMS, which is described in a formal modeling language. Current database-modeling languages include the entity-relationship (ER) model and the unified modeling language (UML). The ER is a conceptual data model that views the real world as entities and relationships. The basic constructs in an ER model are the entities, attributes, and relationships that are in an ER diagram. The ER model focuses on the conceptual and logical design phase of the database. It can be used to develop SQL-compliant database systems, which are convenient for users unfamiliar with database operations (Primer, 2010).

The UML is an object-oriented visual modeling language used to specify, visualize, analyze, and control the objects of a software system. It is used to understand, design, browse, configure, maintain, and control information about software systems (Primer, 2010). This study used the ER model for three important reasons. First, Safety Analyst only supports SQL-compliant databases. Second, ER diagrams, revealing the design of the database, are easier to understand compared to UML diagrams. Third, most applications similar to Safety Analyst are likely to be compatible with an ER model.

The physical data model for the database was built using the ER model, which indicates how data should be represented and stored by a DBMS, such as Oracle, MySQL, SQLServer, or Derby (DynusT, 2008). In this study, the user had the option to choose either MySQL or Derby as a comprehensive database system. However, for the Safety Analyst View, only Derby was enabled because MySQL is not compatible with Safety Analyst. Both databases are open-source, SQL-compliant DBMS, and provide all the required capabilities of a reliable, flexible, and robust DBMS. SQL scripts were developed to generate database tables and the relationships among them in MySQL and Derby.

Once the physical data model for the database was created, the database was ready to be populated with data. Data insertion is a process that can happen once, periodically, or sporadically. The methodology to populate the database was designed to account for most potential scenarios that could arise. For example, various empty tables were designed and created for future data that may become available and/or desirable.

2.4.1 View Tool for Safety Analyst

Such analysis tools as Safety Analyst require data in a particular format. For example, Safety Analyst requires crash severity type in the form of 'K' for fatal, 'A' for severe injury, 'P' for property damage. However, it is unlikely that the data sources use the same formatting. Having the requirement to follow a particular formatting is one of the primary barriers for DOTs to use Safety Analyst (Xiao Qin, 2011). The database developed in this study stores crash severity type in the form of fatal, injury, or property damage.

A view tool for Safety Analyst was developed to provide a database view consistent with the requirements of Safety Analyst. Table 2.3 illustrates a portion of an MS Excel sheet used to

establish mapping between the general database view and the Safety Analyst view. Database Table Name, Attribute Name, and Attribute Values are mapped between the two views. For example, in Table 2.3, the Database Table Name is 'accident,' the Attribute Name is 'severity,' and Attribute Values are 'fatal injury,' 'severe injury,' and 'property damage only.' The corresponding Safety Analyst values are Accident; accidentSeverity1; and K, A, or P.

Table 2.3 Mapping between a General View and the Safety Analyst View

Database Table Name	Attribute Name	Attribute Values	Safety Analyst View Table Name	Safety Analyst View Attribute /name	Safety Analyst View Attribute Value
Accident	severity	Fatal injury	Accident	accidentSeverity1	K
Accident	severity	Severe injury	Accident	accidentSeverity1	A
Accident	severity	Property damage only	Accident	accidentSeverity1	P
RoadwaySegment	routeType	State Route	Roadway Segment	routeType	SR
RoadwaySegment	routeType	Interstate	Roadway Segment	routeType	I
RoadwaySegment	routeType	US Route	Roadway Segment	routeType	US
RoadwaySegment	roadwayClass	Principal arterial - other	Roadway Segment	routeType	3
RoadwaySegment	roadwayClass	Minor arterial	Roadway Segment	roadwayClass	4
RoadwaySegment	roadwayClass	Local	Roadway Segment	roadwayClass	7
RoadwaySegment	roadwayClass	Major Collector	Roadway Segment	roadwayClass	5
RoadwaySegment	roadwayClass	Principal arterial - other freeway or expressway	Roadway Segment	roadwayClass	2
RoadwaySegment	roadwayClass	Principal arterial – interstate	Roadway Segment	roadwayClass	1
RoadwaySegment	roadwayClass	Minor Collector	Roadway Segment	roadwayClass	6
RoadwaySegment	roadwayClass	Other	Roadway Segment	roadwayClass	0
RoadwaySegment	roadwayClass	Unknown	Roadway Segment	roadwayClass	99

The back-end of the view tool for Safety Analyst has a MS Excel parser that streams the data, provides mapping, and stores the data in a matrix. HashMaps are created, and a relationship is established between the database and the Safety Analyst view.

2.5. Analysis and Results

The comprehensive database as well as the database view of Safety Analyst for Clark County, Nevada, was developed with the proposed data management tools, and populated using the data sources described earlier. Using the data management tools, the database view was mapped, imported, and post-processed. Calibration factors for various site subtypes were 1) urban freeway segments with four and six lanes; 2) urban freeway segments in interchange areas with four and six lanes; 3) urban signalized four-leg and three-leg intersections; 4) urban stop-controlled with four-leg and three-leg intersections; and 5) arterial segments with two, four, and six lanes. These factors were obtained by calibrating the federal default SPFs, using Nevada data.

Network screening analysis was performed using the analytical tool in Safety Analyst to determine sites with the most potential for safety improvements. Network screening analysis can be performed using multiple combinations of screening types, safety performance measures, severity, and screening attributes. Results can be reported using two types of reports, 1) conventional, with all the site results; and 2) a percentage type, specifying the percent (e.g., the top 5% sites). Three basic screening types are available that can report 1) the expected and excess crash frequencies, with peak searching on roadway segments using limits for the coefficient of variation (CV) (Hauer, 1997; Wu et al., 2007; ESRI ArcGIS); and 2) a sliding window on roadway segments, and (3) corridor screening (Hauer, 1997; Wu et al., 2007; ESRI ArcGIS). Other screening types analyze a high proportion of specific crash types, a sudden and steady increase in mean frequency, and corridor screening (Wu et al., 2007; ESRI ArcGIS).

Safety performance measures, expected and excess crash frequencies for different severities and various screening attributes can be computed (Hauer, 1997; Wu et al., 2007; ESRI ArcGIS). Using the Clark County database, various analyses using different network screening methods with default SPF (calibration factor =1) and calibrated SPF were conducted for:

- Analysis of roadway and ramp segments and intersections,
- Analysis of roadway segments based on functional classifications,
- Analysis of signalized and stop-controlled intersections, and
- Analysis of ramp segments.

To illustrate the results, this paper reports two case studies that used excess crash frequency as a safety performance measure to see if crashes were reduced if a safety improvement was implemented (Wu et al., 2007). The first case study identified the top 5% sites, including roadway and ramp segments as well as signalized and stops controlled intersections, which have the potential for safety improvements. Two analyses, with default and calibrated SPFs, were performed. Excess crash frequency was calculated for fatal and all injury crashes, with peak searching on roadway segments having coefficient-of-variation (CV) limits for the entire network. The peak-searching screening type was used because it had CV-limit statistics and a minimum window length of a 0.1-mi segment. Hence, the exact section/window of the site that had the potential for safety improvement could be determined to deploy a countermeasure. Seven out of 10 sites were different in the top ranks.

Table 2.4 shows the results of the first case study, including the top 10 sites (the first 10 ranks) having the potential for safety improvements. These sites consisted of two site subtypes,

urban freeway four-lane segments in the interchange area (Site Subtype 158) and urban arterial multi-lane divided segments (Site Subtype 153). Site Subtype 158 had a lower calibration factor, 0.17; implying that these roadways, on average, experienced fewer crashes than roadways used to develop federal SPFs of Safety Analyst. Conversely, Site Subtype 153 had a higher calibration factor, 4.27, implying that these roadways, on average, experienced higher crashes than roadways used to develop federal SPFs. Hence, yearly calibration of SPFs plays a significant role in screening sites that have a higher potential for safety improvements.

Using default and calibrated SPFs, the second case study identified intersection sites with the potential for improvements in both fatal and all injuries. The excess crash frequency for fatal and all injury crashes was calculated. Figures 2.3(a) and 2.3(b) illustrate top 10 intersection sites (the first 10 ranks) having a potential for safety improvements. Two sites (circled in red) with Ranks 4 and 5 vary between analyses with default and calibrated SPFs. This difference is because the sites with ranks three and four have different site subtypes. Hence, different calibration factors were used.

The top 10 sites consisted of two different site subtypes: urban four-leg signalized intersection (Site Subtype 253) and urban three-leg intersection (Site Subtype 254). Site Subtype 253 had a slightly higher calibration factor, 1.08, experiencing higher crashes than the intersections used for developing federal SPFs. Site Subtype 254 had a lower calibration factor, 0.64, implying that urban three-leg signalized intersections experienced less crashes than the intersections used for developing federal SPFs of such sites.

In the results, the predicted crash frequency was much less when compared to the observed crash frequency due to the default SPF in the Safety Analyst. The predicted crash

frequency was one of the important measures for calculating the expected or excess crash frequency by the EB method. Because of the urban nature of the study area, higher levels of AADT (100,000s) worsened these results.

Currently, Safety Analyst calibrates the default coefficients estimated based on national-level data for various site subtypes, such as two-lane freeways, four-lane freeways, using agency AADT data. This issue can be solved in two of the following ways:

a. Create agency-specified site subtypes with different AADT ranges in the administration tool, and recalibrate the coefficients; or

b. Based on the data, develop separate count-regression models for site subtypes, and input the coefficients in the administration tool.

In this study, many case studies were experimented to infer the Safety Analyst results as there are minimum guidelines about a screening type or performance measure to choose for specific analysis. From the inference of results, the following conclusions are obtained:

1) Peak searching screening type was not a good parameter for segments less than 0.1 mi. It proportionated expected/excess crash frequencies for 0.1 mile when the length was less than 0.1 mi. In this case, a sliding window was a better choice because it aggregated and moved the window on contiguous segments for a calculation; further, it proportionated expected/excess crash frequencies for 0.3 mi, the minimum length used to calculate performance measure, when the length of site was less than the window length.
2) In Safety Analyst, peak searching was better as it had the coefficient-of-variation limit, whereas sliding window did not.

Table 2.4 Results of Basic Network Screening with Peak Searching on Roadway Segments and CV tests from Safety Analyst for Fatal and All Injury Crashes on Roadway and Ramp Segments as well as Intersections, using Default and Calibrated SPFs

Analyses Type	Rank	Site Subtype	Route	Location with Highest Potential for Safety Improvement							
				Start Location	End Location	Average Observed Crashes *	Predicted Crash Frequency *	Excess Crash Frequency			
								Excess Frequency *	Variance **	No. of Fatalities	No. of Injuries
Default SPF Calibration Factor (CF) = 1.0	1	158	IR15	40.223	40.323	267.05	21.49	219.41	408.32	2.04	312.21
	2	153	SR589	3.311	3.411	154.14	4.63	139.86	149.66	1.51	211.41
	3	158	IR15	41.386	41.486	173.08	27.10	133.70	609.28	1.24	190.24
	4	158	IR15	35.112	35.768	142.14	23.27	107.04	451.76	0.99	152.32
	5	153	SR612	4.605	5.124	114.01	5.28	102.18	182.06	1.10	154.46
	6	153	SR593	0.889	1.574	103.06	9.51	90.25	544.53	0.97	136.42
	7	153	SR159	29.664	30.193	101.96	8.06	90.19	396.04	0.97	136.33
	8	153	SR612	5.124	5.633	98.84	2.63	86.57	53.97	0.93	130.86
	9	153	SR593	3.784	6.361	94.92	3.99	84.30	107.53	0.91	127.42
	10	153	Las Vegas Blvd	26.032	26.112	87.10	6.22	75.67	239.96	0.82	114.37
Calibrated SPF Site Subtype 158 CF = 0.17 Site Subtype 153 CF = 4.27	1	158	IR15	40.223	40.323	291.56	20.43	238.49	379.89	2.21	339.36
	2	158	IR15	41.386	41.486	189.06	25.76	147.70	558.06	1.37	210.18
	3	153	SR589	3.311	3.411	156.48	19.18	135.02	2165.74	1.46	204.09
	4	158	IR15	35.668	35.768	155.78	22.12	118.56	414.43	1.10	168.71
	5	153	SR612	5.324	5.424	100.37	10.91	87.08	707.86	0.94	131.63
	6	158	IR15	41.567	41.667	118.14	24.19	84.46	483.08	0.78	120.18
	7	153	Decatur Blvd	4.624	4.64	106.95	8.99	77.99	483.52	0.84	117.89
	8	158	IR15	41.667	41.767	105.53	22.24	74.20	409.34	0.69	105.58
	9	153	SR596	5.293	5.393	82.73	7.85	72.04	371.99	0.78	108.89
	10	153	Maryland Pkwy	9.794	9.894	82.69	11.22	69.51	745.94	0.75	105.07

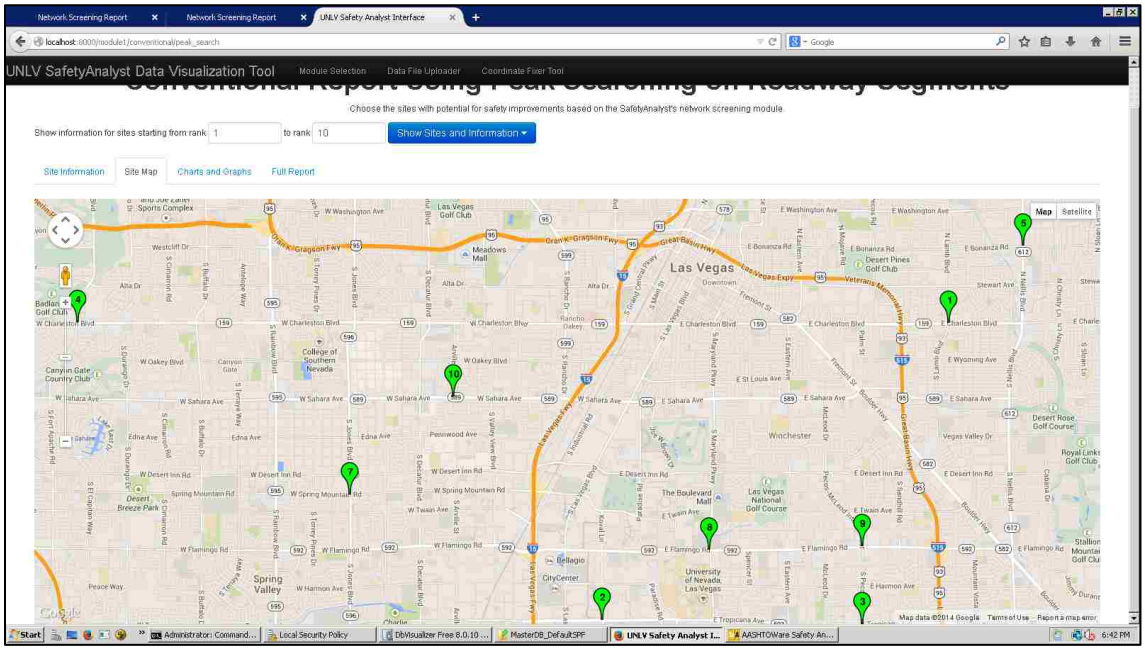


Figure 2.3(a) Results of Basic Network Screening for Fatal and all Injury Crashes at Intersections, using Default SPF.

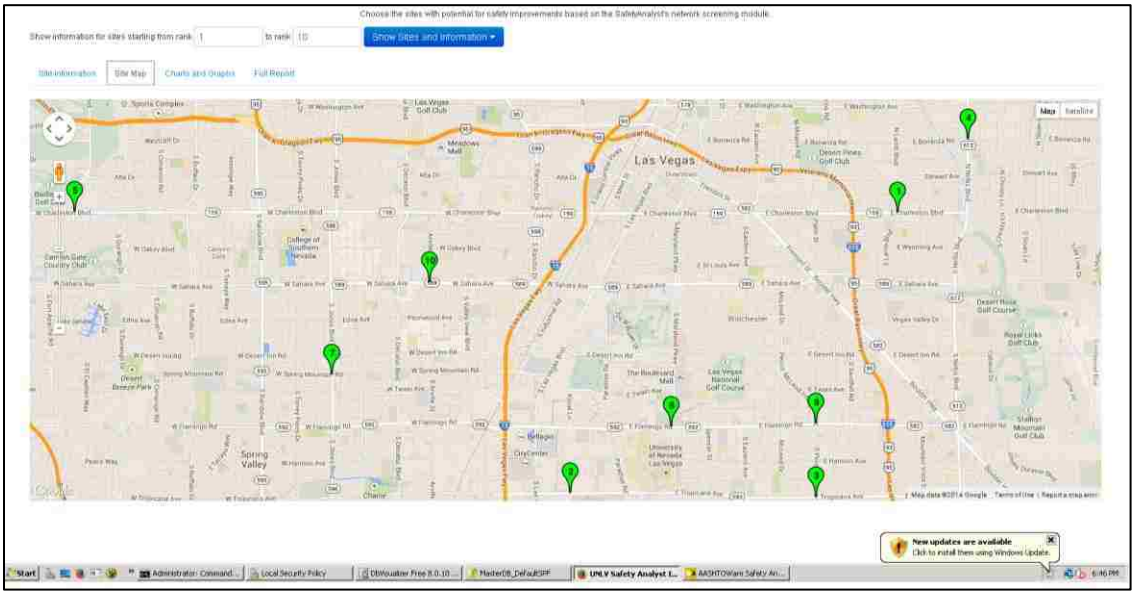


Figure 2.3(b) Results of Basic Network Screening for Fatal and all Injury Crashes at Intersections, using Calibrated SPF.

3) Peak searching was not a good parameter for longer segments. Peak searching provided one rank per site, with a window length of 0.1 mi; other windows with the second highest expected/excess crash frequencies will be provided in additional windows of interest.

Longer segments will have multiple additional windows of interest. However, the sliding window provided consecutive ranks for the same site with various windows.

4) For sites with a higher number of crashes and a large variance, analysts can use either expected or excess crash frequencies.
5) No particular screening type was preferred for the entire analysis. Analysts are recommended to evaluate a given site list using multiple combinations of network screening to find common sites from the output. When the same site is identified using several screening methods, this reinforces that the site deserves further investigation (Wu et al., 2007).

2.6 Visualization Tool for Safety Analyst

The output capabilities provided by Safety Analyst are limited to tables that report the results in HTML, PDF, RTF, and CSV formats. Analysts have to infer the results from these voluminous tables without having an image of the site. Hence, a visualization tool was developed (Song et al., 2007) to provide a better meaning to the output, with expanded capabilities for spatial, graphical, and editable reports. To visualize the results using the visualization tool, the user can choose between two alternative display methods: Google Maps and ArcGIS. The advantage of using Google Maps is its simplicity and availability; the advantage of ArcGIS is its modeling and computing capabilities.

For the Google Map interface, the visualization tool has a web-based front-end; for the ArcGIS interface, the visualization tool is a standalone application based on ArcPy scripts. Both applications provide easy access to multiple tabs. The tabular results and a map with spatial locations are displayed in the first two tabs. In addition, the user can interact with the graphical display to perform such basic operations as zoom-in, zoom-out, and select sites. In the second

tab, the user can choose the bar chart for the safety performance measures, such as observed, predicted, and expected/excess crash frequencies for several sites.

Interpreting the results by means of graphs is easier than by tables. The user can include site spatial locations and performance measure bar charts in the editable Safety Analyst report in the third tab. The tabular results of network screening provided in Table 4 are difficult to use without the visualization tool. However, Figures 2.3 (a) and (b) illustrate the spatial locations of intersections by using the developed visualization tool (Song et al., 2007), determined through network screening.

2.7 Conclusions

The benefits of developing and using a comprehensive database system for traffic safety studies are significant. This study developed a comprehensive database system that can provide data to multiple applications for traffic safety engineering and other potential needs. Furthermore, it provided the methodology and guidance to develop a database from the existing, readily available data sources at the state DOTs and/or MPOs. In addition, the tools developed to build the comprehensive database and view for Safety Analyst can be used by other agencies, as they use non-commercial software. This system allows the use of state-of-the-art traffic safety tools to support the development of federal requirements as well as develop better traffic safety solutions for existing and emerging problems. These tools offer significant savings in terms of time, money, and lives.

In particular, the proposed database system has the capability to provide data to Safety Analyst, the state-of-the-art highway safety management software. Although Safety Analyst provides tremendous analysis capabilities, few agencies take advantage of these capabilities

because the software requires significant data needs, complex development of the required inputs, and lack of experience and knowledge in creating the inputs as well as using the software (Xiao Qin, 2011). The proposed database system, along with its data management and visualization tools, provides significant support to circumvent these barriers. This database system can be used to develop jurisdiction-specific SPFs for better estimation of performance measures.

This study can be expanded to develop tools that create different site subtypes based on the data in the Safety Analyst view. SPFs can be developed for those site subtypes, and the coefficients can be inputted into the administration tool to obtain better predictions for crash frequency. Predictive methods in Part C of the *Highway Safety Manual* (Alluri, 2008) can be used for SPF development. In this case, the developers of Safety Analyst should expand the capabilities of an administration tool to accommodate the agency-specific multi-parameter SPFs and their coefficients.

CHAPTER 3

A BUSINESS INTELLIGENCE FRAMEWORK FOR TRAFFIC SAFETY NETWORK SCREENING

3.1 Introduction

The significance of ensuring traffic safety is the focus of such federal legislation as the Transportation Equity Act for the 21st Century (TEA-21), the Safe Accountable Flexible and Efficient Transportation Equity Act - A Legacy for Users (SAFETEA-LU), and the Moving Ahead for Progress in the 21st Century (MAP-21). SAFETEA-LU and MAP-21 both require that states develop comprehensive Highway Safety Improvement Plans (HSIPs) (FHWA, 2013). One of the critical programs of HSIPs is the traffic safety management process, which involves annual reporting of the highway locations that exhibit the most severe traffic safety needs. By identifying the most hazardous roadway site locations, specific countermeasures can be implemented that would improve safety conditions. In a traffic safety management process, identifying locations with the potential for safety improvements is known as *network screening*, as described in Part B of the *Highway Safety Manual* (*HSM*) (AASHTO, 2010).

Despite the availability of sound methodologies as expected/excess crash frequency by an empirical Bayes (EB) method (recommended by the *HSM*), practitioners continue using theoretically unsound methodologies which rely only on observed crash frequency or crash rates for network screening. For HSIP reports submitted to the Federal Highway Administration in fiscal year 2014, only four states used theoretically sound methods for network screening as described in the *HSM* (FHWA, 2015). Barriers that prevent the use of theoretically sound methods include: 1) significant data needs and integration are required, 2) a special schema is

required to enable analysis using specialized software, 3) time-consuming and intensive processes are required, 4) relevant technical knowledge is lacking, 5) visualization capabilities are lacking, and 6) coordination across various data owners is required.

Alluri and Oogle (2012) documented the current safety-analysis practices related to engineering as used by various states; in addition, they described perspectives in adopting and implementing the methods provided in the *HSM*. They indicated that barriers faced by traffic safety engineers include requirements for comprehensive data sets, data integration, and management. Tarko et al. (2014) and Paz et al. (2015c) discussed the complexities of data integration and management for network-level traffic safety analysis, specifically for the traffic safety management process.

This paper proposes a framework to address the above listed barriers to enable practitioners to use theoretically sound methodologies for network screening. The framework was developed using a Business Intelligence (BI) approach, which provides methods and mechanisms to integrate and process data, generate advanced analytics, and visualize results. The proposed framework facilitates traffic safety engineering and enhances the outcomes of an HSIP.

In transportation engineering and traffic safety, several approaches have been developed. GIS methodologies developed by ESRI® have been widely used for data processing, analysis, reporting, and visualization (Pulugurtha et al., 2007; Wellner Qin, 2011; Aylo 2010). The Critical Analysis Reporting Environment (CARE), developed by the University of Alabama (CAPS, 2009a), sorts, analyzes, and compares crash data using functions that allow statistical analyses with charts and graphical displays (Paz et al., 2014, Khanal, 2014). For comprehensive traffic safety management, Safety™ Analyst is a state-of-the-art software (AASHTOWare™) that was developed using network screening methods from the *HCM* (AASHTO, 2011). Although Safety

Analyst provides significant capabilities, the software has several limitations. For example, data integration and processing capabilities are lacking, the data needs to be in a specific schema, and it has marginal visualization capabilities. Tjandra (2014) developed a BI system for traffic data integration by linking roadway, crash, and traffic flow data to improve traffic safety. This system provides descriptive performance indices for traffic safety.

As yet, no single framework exists that provides capabilities for 1) data process, integration and management, 2) advanced analysis, and 3) visualization. Key characteristics of the BI framework proposed in this paper include: 1) an extract-load-transform (ELT) process; 2) tools for integration of data from a wide variety of data sources; 3) algorithms for theoretically sound analysis, as recommended by the HSM; 4) a methodology for effective corridor-level network screening; and 4) visualization tools for network-wide site-specific and corridor-level analysis. These characteristics provide an effective platform for generating theoretically sound analysis and information for various types of decision makers. Furthermore, as new source data is provided to the proposed framework, all analyses, reports, and visuals are updated.

Currently, for network screening for individual sites (roadway segments, ramps, and intersections), the *HSM* recommeds using expected/excess crash frequency by an EB method. Corridor-level network screening is important for decision makers because it enables to rank corridors rather than sites so as to provide homogenous infrastructure to minimize changes within relatively short distances. Improvements are recommended for long sections of roadways that could include multiple sites with the potential for safety improvements. Few agencies have corridor-wide safety programs. Some program that are in place include Nevada's Kietzke Lane Safety Management Plan; the Safe Corridor Programs of both New Jersey and Wisconsin; and

Integrated Corridor Management plans that develop safety plans/programs for cities and municipalities (Shimko, & Walbaum, 2010; Qin et al., 2013).

Corridor-level network screening is important for such programs to identify corridors that have safety needs. Several studies have used observed crash frequency, crash rates, or a crash severity index for corridor screening (AASHTO, 2011; Hamidi et al., 2010; Qin et al., 2013). Using observed crash frequencies result in a volume bias, while using crash rates result in a segment length bias; in addition, using observed crashes result in a regression-to-the-mean bias. For corridor screening, Hamidi et al. (2015) used crashes that occurred only on major roads at intersections. Ignoring interactions of major and minor road characteristics at intersections affects predicted crash frequency, leading to incorrect estimation of expected crash frequency. Other studies did not search corridors that had potential for safety improvements, but instead estimated the crash frequency on pre-aggregated sites. (AASHTO, 2011; Zhao et al., 2014). As an alternative for defining corridors for implementing safety improvements based on pre-aggregated sites or lenghts, determining them using a sliding window mechanism based on characteristics and crash data provides a superior approach. A sliding window mechanism addresses crash location errors by evaluating the same section of roadway multiple times, using overlapping windows.

The contributions of this research include a comprehensive framework for network screening, using concepts in data warehousing and Business Intelligence as well as a methodology for corridor-level network screening. Required data sources include those commonly used by state agencies.

3.2 Methodology

To illustrate the advantages of the proposed framework, network screening algorithms from the *Highway Safety Manual* were reimplemented and expanded using the Oracle® Business

Intelligence Enterprise Edition (OBIEE) (Rittman, 2013). Oracle Data Integrator (ODI) (Dupupet et al., 2013) was used to develop a safety data warehouse, which was accessed by OBIEE to facilitate the development of advanced analytics, dashboards, and maps. The connection to the database was created by the Repository Design Model (RPD), which contains physical models, business mapping models, and presentation models for use by OBIEE (Rittman, 2013). Oracle R Enterprise (McDermid & Taft, 2014) scripts were developed to implement network screening algorithms. These scripts were executed in the physical layer (Rittman, 2013) of the RPD.

The output of the Oracle R Enterprise scripts was saved in datastores, which allowed other queries to access the results for network screening. These queries were used in the RPD to enable OBIEE to perform on-the-fly computation and retrieval of the network screening results in the dashboards. The JavaScript application program interface (API) (ArcGIS Web API, 2015) for Esri® maps was used in the dashboards along with analytics to display network screening results and associated site locations spatially.

3.2.1 Data Warehouse Design with ODI

Silos of source data from various sources can be integrated with ODI to create a safety data warehouse. Source data includes the road network, traffic volumes, and the Highway Performance Management System (HPMS) as well as crash data and their associated characteristics. The data warehouse was developed using an ELT process. ODI interfaces extract data from the source, and loads the information, using a Loading Knowledge Module (LKM), into the OBIEE target database (Rittman, 2013). In this study, the data was transformed into a star schema for use in OBIEE.

Data across systems were integrated using a location reference system, County/Route/Milepost. In this study, crashes and their characteristics were mapped to the

segments, intersections, and ramps. Similarly, traffic stations were mapped to road segments in order to obtain the average annual daily traffic (AADT) on respective road segments. Physical characteristics of road segments – including the number of lanes, median type, median width, speed limit, operation-way, area type, and access control – were obtained from HPMS.

Data from signalized intersections can be obtained from such sources as the Freeway and Arterial System of Transportation (FAST) (Xie & Hoeft, 2014) of the Regional Transportation Commission of Southern Nevada (RTC-SN). Data from stop-controlled intersections can be collected using Google Earth. ODI can be used to integrate the data from intersections with data from the road network as well as with crash data. Figure 3.1 illustrates an ELT process for crash-related information from various tables of crash data to a target database table, SA_ACCIDENT. Similarly, three target database tables were created, SA_ROADWAYSEGMENT, SA_INTERSECTION, and SA_RAMP.

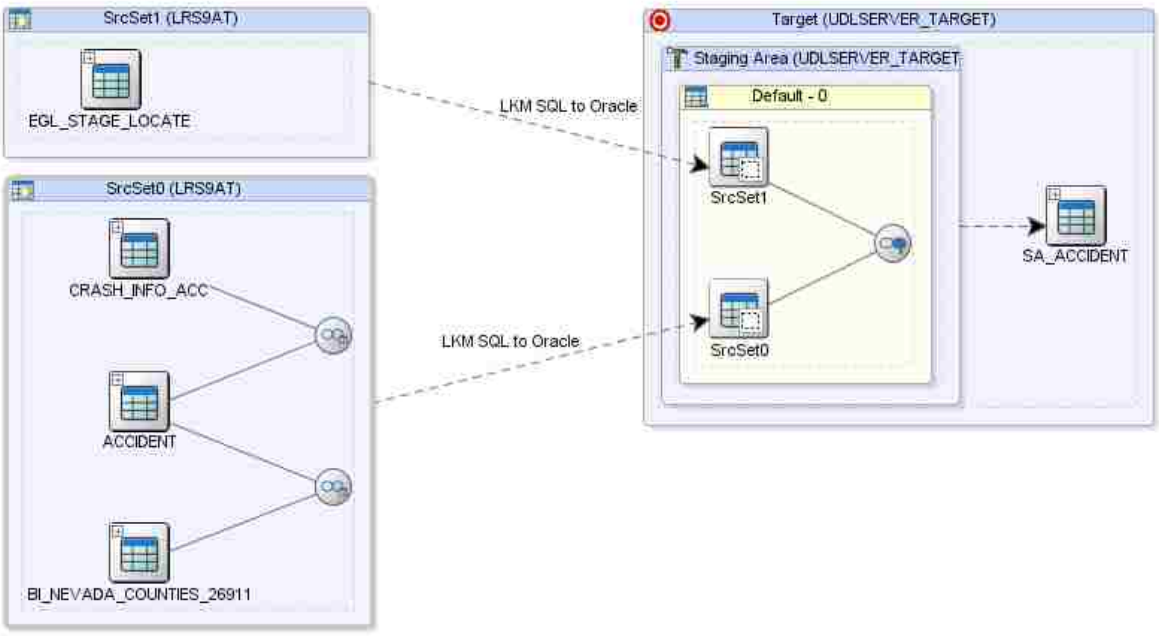


Figure 3.1 ELT Process of Crash Information to SA_ACCIDENT Target Table.

Contiguous sites with similar physical characteristics needed to be aggregated in order to create homogeneous segments. Procedures were developed using Oracle Procedural Language/Structured Query Language (Pl/SQL), and were connected to a web-based interface for homogeneous segmentation. As a result, analysts and engineers could use a Choice List in the interface to choose parameters that can be used for homogeneous segmentation. These include such parameters as a district, county, or route; the number of through lanes in one direction or a combined direction; median type; median width; and the percentage of the AADT threshold. By using this interface, as shown in Figure 3.2, the segments can be aggregated, and the new site list created and stored in the target database for further analysis.

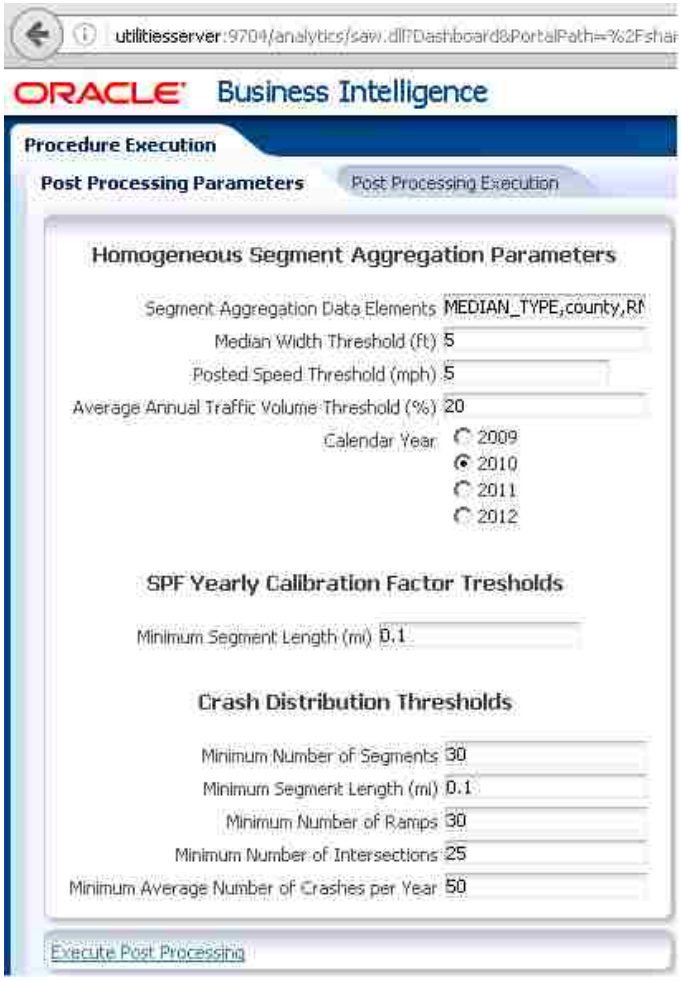


Figure 3.2 Dashboard interface for post processing and calibration.

Once the site list is created, the sites with characteristics for area code, functional class, number of lanes, access control, and median type can be used to group sites into site subtypes. This operation easily can be performed using a single SQL statement. Sites with the same site subtypes are used to estimate predicted crash frequency. Predicted crash frequency is estimated using a calibration factor multiplied with the safety performance function (SPF), as documented in the *HSM* (AASHTO, 2010). National default values for safety performance functions can be obtained from the *HSM* for all site subtypes. Calibration factors can be calculated as the ratio of the sum of observed crash counts from the target database to the sum of the predicted crash counts from the safety performance function.

All the procedures mentioned above were implemented in ODI, and tables were created to store results in the target database.

3.2.2 Network Screening Using Oracle R in RPD

Network screening is a systematic review process that identifies and ranks roadway sites for potential safety improvements. This process is critical because a detailed engineering study for all network sites is expensive. The purpose of network screening is to review the entire roadway network, or portions of the roadway network, and identify and prioritize sites with promise for safety improvements. These identified sites are recommended for further investigation and a detailed engineering study.

The three network screening algorithms used in this study were 1) Peak Search, 2) Sliding Window, and 3) Simple Ranking. The first two were used for roadway segments, and the third one was used for intersections and ramps; the second one was used for corridor-level network screening as well. The *HSM* and other literature (AASHTO, 2010; Paz et al., 2014) identified Peak Search

and Sliding Window as the two recommended algorithms for performing network screening along roadway segments.

To use the Peak Search algorithm, the roadway segment of interest is divided into windows of equivalent length that do not overlap; then, a performance measure of interest is calculated. A small window length of 0.1 mi is evaluated first, and is adjusted gradually for greater lengths. The coefficient of variation (CV) is calculated for each segment, which is the ratio of the standard deviation to the mean of the expected value. If the standard deviation is less than the mean of the expected or excess crash frequency (i.e., a small CV value), this indicates a high level of precision in the estimate. Thus, a smaller CV increases the user confidence level regarding the results, and vice versa (AASHTO, 2010).

In the Sliding Window algorithm, the user selects a pre-defined window length. The algorithm estimates the performance measure for the window, and then slides the window by incremental lengths to estimate the performance measures of the subsequent windows. All the windows are ranked with regard to the estimated performance measure.

In contrast to the Peak Search and Sliding Window methods, the Simple Ranking approach is used when considering roadway components, such as intersections or ramps, as a single entity. These components are ranked using the estimated performance measures. Details of the algorithms for all network-screening methods can be obtained from Part B of the *HSM* (AASHTO, 2010).

In this study, network-screening algorithms were developed using Oracle R scripts. OBIEE use Oracle R scripts to execute the network screening algorithms. These R scripts were saved to the database by using Oracle R Enterprise libraries, and can be executed with the 'rqTableEval' stored procedure (McDermid & Taft, 2014). An R script that has a final data frame to return will output a standard Oracle database table when executed.

Two sets of R functions are saved in the Oracle database. One set of R scripts responsible for performing the network screening, getting results, and saving the results as a data frame to a datastore, which is a table accessible with the Oracle R Enterprise libraries that allows R variables to be saved to the database. The second set of scripts is responsible for loading the data frame (McDermid & Taft, 2014) from the datastore and returning the data frame. An Oracle SQL select statement can be used to execute these R scripts. By saving the SQL select statement as a view and loading the view into the physical layer of the RPD, OBIEE is able to execute network screening and load the results (Rittman, 2013).

The data required for network screening algorithms are accessed from the target database mentioned in Section 3.2.1. The view with the results is called a fact table and the target database tables are called dimension tables (Rittman, 2013). The star schema (Rittman, 2013) created in the physical layer of the RPD is illustrated in Figure 3.3. These layers are brought into the business layer and the presentation layer for further analytics. The business layer performed joins, which helps mapping site locations in Fact table as well as crash, roadway, ramp and intersection characteristics in their respective dimension tables (Rittman, 2013).

3.2.3 Corridor Screening

Corridor-level network screening provides the capability to compare the safety performance across extended corridors rather than comparing the safety performance of individual sites (AASHTO, 2010). A corridor may consist of multiple such sites as roadway segments, intersections, and/or ramps, which are combined together to analyze as a single entity.

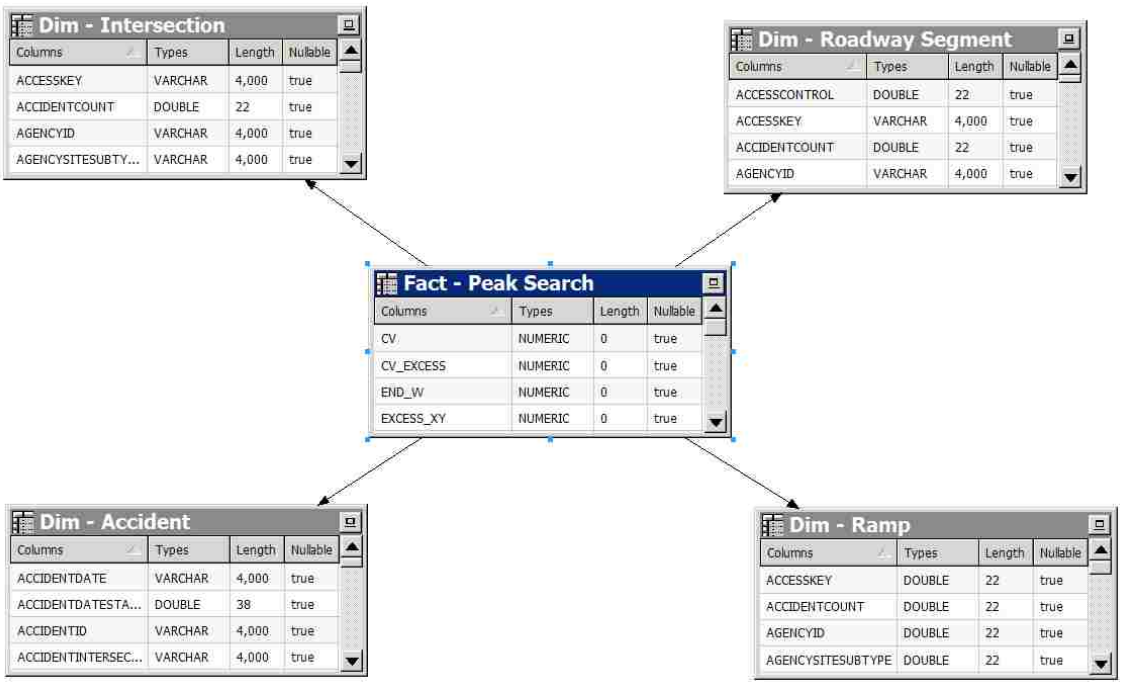


Figure 3.3 STAR Schema for the Peak Search Network Screening.

In this study, two types of corridor-level network screening algorithms are proposed: 1) Fixed Corridor screening, and 2) Corridor Search. Fixed Corridor screening can be used for pre-defined corridors. When a user specifies the predefined corridors, the expected crash frequency of these corridors is estimated by aggregating the expected crash frequencies of individual roadway elements. These predefined corridors are ranked from the highest to the lowest with regard to expected crash frequencies. This method is useful when engineers are evaluating known corridors in the network. Sites are assigned by the engineer/analyst to a specific corridor in a specific table at the data management level. If sites assigned to the corridor need to be modified or more corridors need to be added, the analyst is required to perform this operation at the data management level.

Corridor Search reviews a road network in a systematic manner to identify the corridors, using a corridor length and an incremental length. The user selects a predefined length to estimate

the expected crash frequency of the corridor and also selects a predefined incremental length that slides the corridor to evaluate next corridor. For each corridor, the expected crash frequency is estimated by aggregating the expected crash frequencies of individual roadway elements, such as roadway segments, intersections, and/or ramps. Then, corridors in the network are ranked from highest to lowest with regard to expected crash frequencies. Moving the corridor by a small incremental length is used to compensate for sites being falsely selected that had randomly high crash counts.

The methodology for estimating the performance measure Expected Crash Frequency for corridors, considering total crashes, is provided in the following steps, based on the *HCM* (*AASHTO, 2010*). Notations and their descriptions used in the equations are:

PF_{yi}	Predicted crash frequency of site *i* in year *y*
SPF_{yi}	Safety performance function of site *i* with data corresponding to year *y*
α, β_1 *and* β_2	Estimated model parameters using safety performance function
c_{yi}	Calibration factor for site *i* in year *y*
C_{yi}	Yearly correction factor for year y relative to year 1 at site *i*
O_{yi}	Number of observed crashes for year *y* at site *i*
w_i	Weights calculated for EB method
b_i	Over-dispersion parameter obtained from SPF regression for site *i* belonging to corresponding site subtype
L_i	Length of site *i*
EF_i	Expected crash frequency using EB method of site *i*
EFc	Expected crash frequency of the considered corridor

Step 1: Calculate the predicted crash frequency per mile for roadway segments and intersections and ramps in a corridor for each year using Equations 3.1 and 3.2, respectively. Usually, the data contains various site subtypes of roadway elements. Hence, appropriate safety performance function (SPF) model parameters, α and β, for associated site subtype needs to be used. A SPF for site subtypes estimated using local data is preferred over one available in the literature. SPFs estimated using data from other regions need to be calibrated using local information. A calibration factor multiplies the SPFs.

$$PF_{yi} = SPF_{yi} = c_{yi} * e^{\alpha} * AADT_{yi}^{\beta} \quad (3.1)$$

$$PF_{yi} = SPF_{yi} = c_{yi} * e^{\alpha} * AADT_{yi}^{\beta 1} * AADT_{yi}^{\beta 2} \quad (3.2)$$

Step 2: Compute yearly correction factors for number of years considered in the data using Equation 3.3.

$$C_{yi} = \frac{PF_{yi}}{PF_{1i}} \quad (3.3)$$

Step 3: Compute weights, w, to be used in empirical-bayes method to provide weightage for observed and predicted crashes using Equation 3.4.

$$w_i = \frac{1}{1+b_i \sum_{y=1}^{Y} PF_{yi} * L_i} \quad (3.4)$$

Step 4: Calculate expected crash frequency for the first year of data for the site *i* using Equation 3.5. The unit of expected crash frequency is crashes per mile per year. In the case of intersections and ramps, the length is '1' and the units are crashes per year.

$$EF_{1i} = w_i PF_{yi} + \frac{(1-w_i) \sum_{y=1}^{Y} O_{yi}}{L \sum_{y=1}^{Y} C_{yi}} \quad (3.5)$$

Step 5: Calculate expected crash frequency for the final year of data for site *i* using Equation 3.6. The unit of expected crash frequency is crashes per mile per year. As in Equation 3.4, in the case of intersections and ramps, the length is '1' and the units are crashes per year.

$$EF_{Yi} = EF_{1i} * C_{Yi} \quad (3.6)$$

Step 6: Calculate expected crash frequency for the entire corridor using Equation 3.7.

$$EF_C = \sum_{i=1}^{I} EF_{Yi} \quad (3.7)$$

Note: Note: For corridor search using the sliding corridor mechanism, starting and ending sites in the corridor could be a fraction of a site. For these cases, the length of the site, *L*, is the length of a fraction of the site considered in the corridor. Similarly, for observed number of crashes, the number of crashes on the corresponding fraction of site should be used.

Both Fixed Corridor and Corridor Search algorithms, including ELT process and star schemas, were implemented in the proposed BI framework.

3.3 Results and Discussion

3.3.1 Results of Data Management

An interface was created using an OBIEE dashboard to execute the developed procedures for homogeneous segmentation of roadway segments. In Figure 3.2, as shown in Section 2.1, an input section for parameters used for homogeneous segment aggregation was shown, by which the user could enter aggregation data elements and threshold values for median width, posted speed, and AADT. Based on the parameters entered, the aggregation of roadway segments is performed. In addition, the user could perform calibration and crash distribution using the same post processing interface.

Minimum segment length for calibration and threshold inputs for crash distribution were provided in order to execute the post processing, using the link, Execute Post Processing, as shown earlier in Figure 3.2. In this study, homogeneous segment aggregation was performed using following parameters 1) number of through lanes, 2) median type, 3) 20% AADT threshold, 4) five ft. median width threshold, and 5) five mph posted speed limit threshold. Minimum segment length of 0.1 mi was used for calibration. Once the post processing was performed, the results were saved as database tables. Various post processed tables were created, including 1) homogeneously segmented datasets for roadway segments, 2) intersection and ramp dataset tables with associated site subtypes, 3) tables with calibrated factors for site subtypes, and 4) tables with crash distribution values for all crash types. Later, these tables were accessed by R scripts to perform network screening.

3.3.2 Results of Network Screening for Peak Search and Sliding Window

A web interface was designed and implemented to run network screening on the fly, using OBIEE Presentation Services (Rittman, 2013). This included a dashboard prompt for parameter inputs, analytics for the presentation of performance measures and other related information as well as filters for specific values to activate dashboard prompts.

A dashboard with the dashboard prompt was created using the presentation variables for input parameters, as shown in Figure 3.4. As a first step, a user has to select the network screening algorithm. Then, a section would be expanded with the dashboard prompt that has radio buttons to input crash severity variables (CSV), screening performance measures (SPM), type of screening (Type), CV threshold, and the limiting performance measure (*XY* threshold) for flagging sites. With this user input, the analyst can screen the network for various crash (collision) types or can select particular days of week or months. In addition, the name of the

analysis can be provided, which enables multi-user analyses. With this functionality, various users can perform analyses and display results on the dashboard, based on the analysis name. Once the input parameters are entered, user has to click 'Apply' to set a platform for the type of the analysis. The 'Run Network Screening' link enables running the analysis. The 'View Network Screening Results' link provides a view of the results.

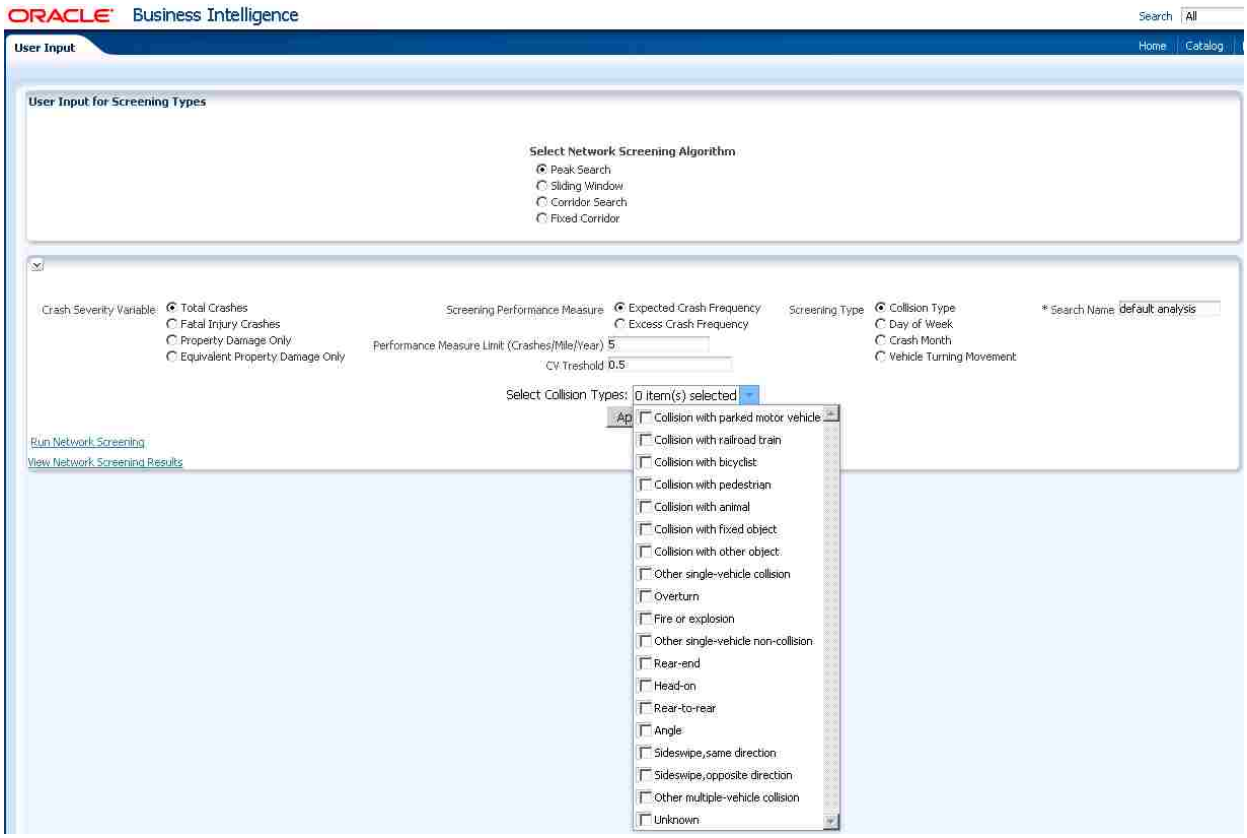


Figure 3.4 Dashboard Illustrating the User Input Interface for Network Screening.

Analytics were created using columns from two tables, Fact-Peak Search and Dim-Roadway Segment. The columns used in this study to present the results were Route Name, Agency ID, Site ID, Window Begin, Window End, Expected Crash Frequency, Excess Crash

Frequency, and Variance. Filters were created to filter ranks, and the name of the analysis. The dashboard was created using the various analysis objects, including tables, graphs, and maps.

Users have an option of selecting an analysis name with a drop-down menu as well as ranks, using the analysis prompt in the analysis section. An Esri map was created in the dashboard to present the spatial location of roadway segments, which are color-coded based on their ranks. Selecting the ranks in the analysis would filter the segments in the map. Figure 3.5 shows a snapshot of a Peak Search analysis using data from Nevada.

Users can drill down further on the segment to diagnose a high-crash location for detailed characteristics of crashes and roadway segments. These characteristics provide the crash pattern in the site location, such as a high number of night-time rear-end crashes. In addition, the user can turn on a Google Earth satellite image for further site information. This information may provide insights to the user to determine countermeasures that could mitigate crashes.

A snapshot of a drill-down analysis is shown in Figure 3.6. The figure shows a description of crash severity and crash (collision) types for a top-ranked roadway segment from a screening analysis. Distributions for light conditions, crash time of day, day of week, number of vehicles, vehicle types involved, and weather condition can be created and displayed in the same drill-down analysis. These distributions provide a clear picture in order to select the type of countermeasure to mitigate future crashes. The user can export the analysis to a portable document format, Microsoft Excel, or PowerPoint by using the export or print tools inherent in OBIEE. This information can be disseminated to decision makers by means of email.

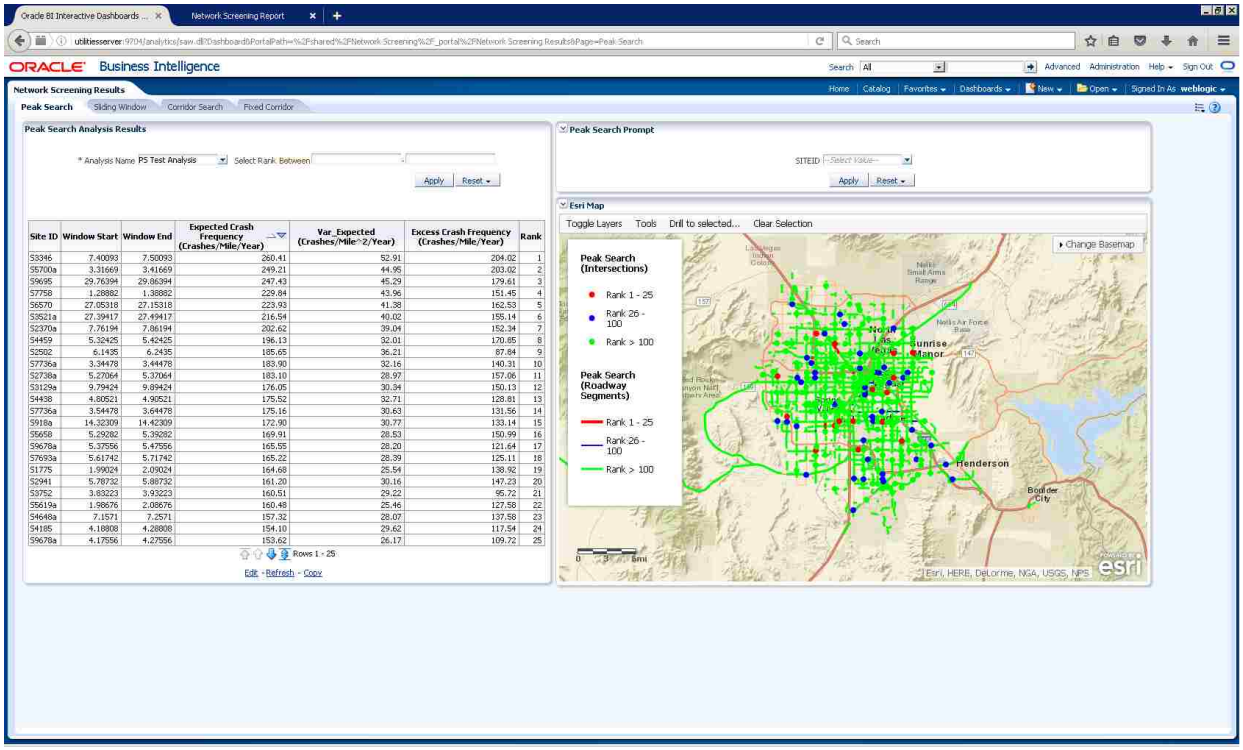


Figure 3.5 Dashboard Illustrating Results and Visualization of Peak Search Network Screening.

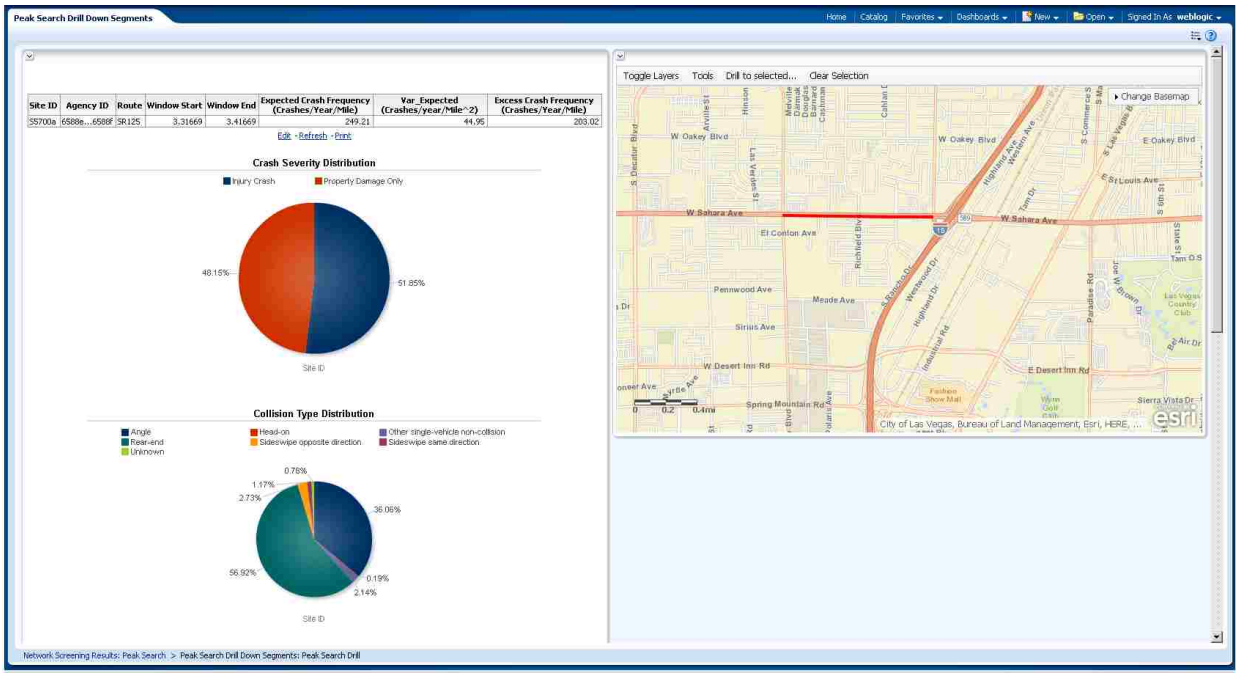


Figure 3.6 Dashboard Illustrating Drill Down Results of a Roadway Segment Results.

Similar to Peak Search, the Sliding Window dashboard was designed and implemented to run analyses and display results; the only difference is the input parameters. The results of the analysis are stored in the View, as discussed in Section 3.2.2. Analytics created using the table, Sliding Window View, display the results for expected and excess crash frequencies in the dashboard, using tables and maps. The drill-down analysis was created to diagnose characteristics for crashes, roadways, and traffic at high-crash locations.

3.3.3 Results from Corridor Screening

Both the Fixed Corridor and Corridor Search algorithms were implemented for network-wide corridor screening. Dashboards were prepared using the same concept to screen or search the corridors.

3.3.4 Results from Fixed Corridor Screening

For Fixed Corridor screening, corridors were predefined in the roadway segment dataset. For illustration purposes, approximately five miles of corridors were predefined and analyzed. An Esri map was created in the dashboard to display the spatial locations of fixed corridor results. Based on route, begin and end mile of predefined corridors, the geometry of segments within a corridor is displayed in the ESRI map. The results of top 10 fixed corridors that are displayed on the dashboard are shown Figure 3.7. Results include the corridor ID, sites in the corridor, route, begin mile, end mile, the expected crash frequency, and the ranks of the corridor. In the results, the column ‘Sites in the Corridor’ includes sites that contain roadway segments, intersections and ramps in the corridor. Ranks of the fixed corridors are based on the expected crash frequency of the corridor.

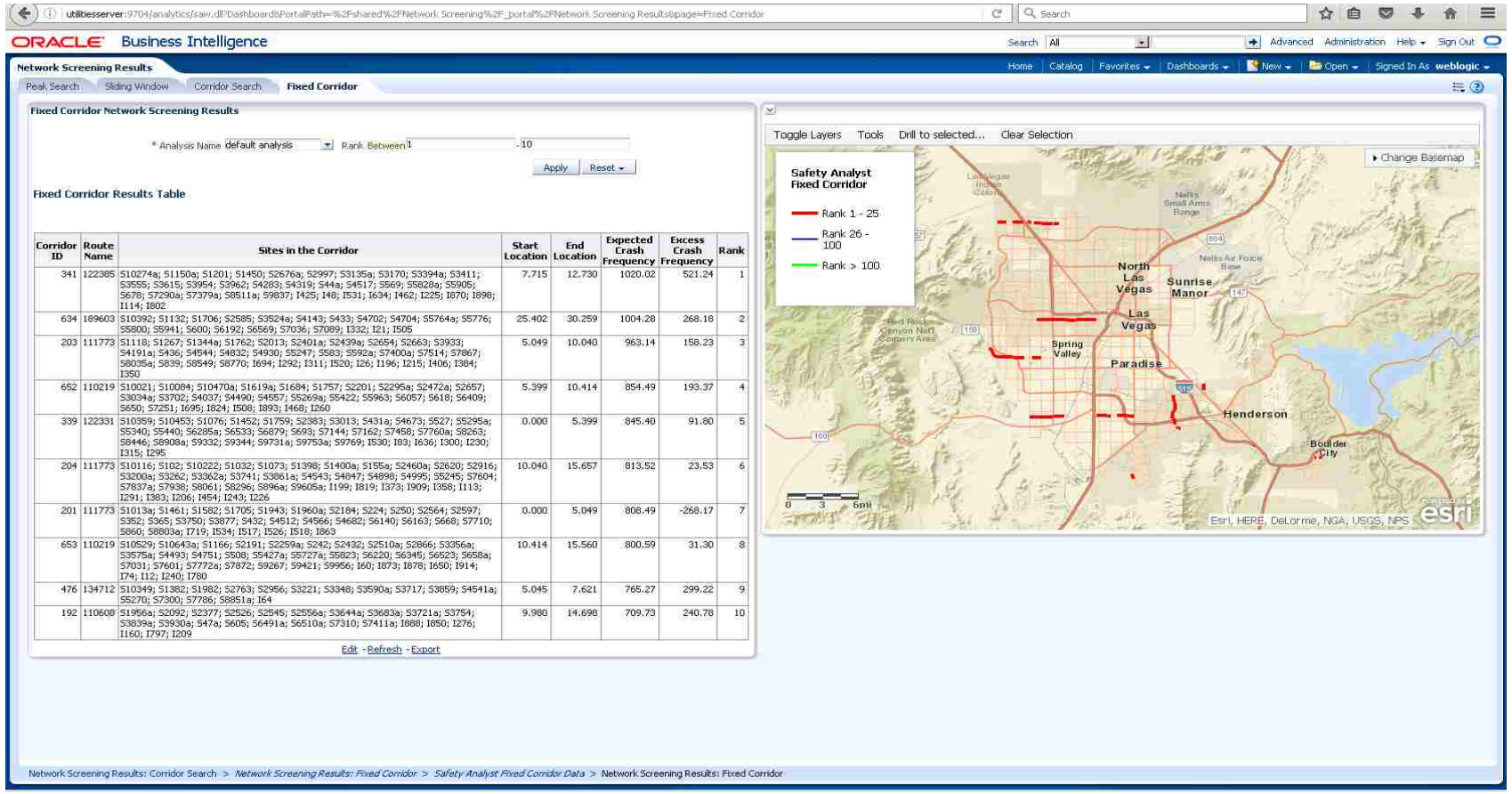
The results obtained in this study were compared with the results using AASHTOWare Safety Analyst. The current literature for fixed-corridor screening uses crash frequency and rate

methods, which also are used by this software. The literature states that when considering extended corridors for analyses, there is less variability or randomness in the crash data. Hence, frequency and rate methods provide accurate corridors for safety improvements. Fixed Corridor screening was analyzed using Safety Analyst for the same corridors used in this study.

Table 3.1 shows the top 15 ranked corridors obtained in this study, using the EB adjusted expected crash frequency, and the corresponding corridor ranks using the observed crash frequency and crash rate methods from Safety Analyst. The ranks obtained using all the three methods were different due to the use of safety performance functions and the advantages of EB-adjusted expected crash frequency that were incorporated in this approach.

Table 3.1 Comparison of Ranks of Top 15 Fixed Corridors using EB Expected Crash Frequency, Observed Crash Frequency and Crash Rate Methods

Corridor Route	Corridor Begin Mile	Corridor End Mile	Rank - EB Expected Crash Frequency	Rank - Observed Crash Frequency	Rank - Crash Rate
122385	7.715	12.73	1	14	4
189603	25.402	30.259	2	21	24
111773	5.049	10.04	3	10	13
110219	5.399	10.414	4	12	11
122331	0.000	5.399	5	19	18
111773	10.04	15.657	6	9	7
111773	0.000	5.049	7	16	22
110219	10.414	15.56	8	11	11
134712	5.045	7.621	9	8	14
110608	9.98	14.698	10	20	23
128	0.000	3.784	11	2	19
137611	5.067	7.924	12	4	8
129234	5.324	10.34	13	17	9
119961	0.472	5.569	14	15	3
113550	5.016	8.416	15	7	15



Corridor ID	Route Name	Sites in the Corridor	Start Location	End Location	Expected Crash Frequency	Excess Crash Frequency	Rank
341	122385	S10274a; S1150a; S1201; S1450; S2676a; S2997; S3135a; S3170; S3394a; S3411; S3555; S3615; S3954; S3962; S4283; S4319; S44a; S4517; S569; S5828a; S5905; S678; S7290a; S7379a; S8511a; S9837; I425; I48; I531; I634; I462; I225; I870; I898; I114; I802	7.715	12.730	1020.02	521.24	1
634	189603	S10392; S1132; S1706; S2585; S3524a; S4143; S433; S4702; S4704; S5764a; S5776; S5800; S5941; S600; S6192; S6569; S7036; S7089; I332; I21; I505	25.402	30.259	1004.28	268.18	2
203	111773	S1118; S1267; S1344a; S1762; S2013; S2401a; S2439a; S2654; S2663; S3933; S4191a; S436; S4544; S4832; S4930; S5247; S583; S592a; S7400a; S7514; S7867; S8035a; S839; S8549; S8770; I694; I292; I311; I520; I26; I196; I215; I406; I384; I350	5.049	10.040	963.14	158.23	3
652	110219	S10021; S10084; S10470a; S1619a; S1684; S1757; S2201; S2295a; S2472a; S2657; S3034a; S3702; S4037; S4490; S4557; S5269a; S5422; S5963; S6057; S618; S6409; S650; S7251; I695; I824; I508; I893; I468; I260	5.399	10.414	854.49	193.37	4
339	122331	S10359; S10453; S1076; S1452; S1759; S2383; S3013; S431a; S4673; S527; S5295a; S5340; S5440; S6285a; S6533; S6879; S693; S7144; S7162; S7458; S7760a; S8263; S8446; S8908a; S9332; S9344; S9731a; S9753a; S9769; I530; I83; I636; I300; I230; I315; I295	0.000	5.399	845.40	91.80	5
204	111773	S10116; S102; S10222; S1032; S1073; S1398; S1400a; S155a; S2460a; S2620; S2916; S3200a; S3262; S3362a; S3741; S3861a; S4543; S4847; S4898; S4995; S5245; S7604; S7837a; S7938; S8061; S8296; S896a; S9605a; I199; I819; I373; I909; I358; I113; I291; I383; I206; I454; I243; I226	10.040	15.657	813.52	23.53	6
201	111773	S1013a; S1461; S1582; S1705; S1943; S1960a; S2184; S224; S250; S2564; S2597; S352; S365; S3750; S3877; S432; S4512; S4566; S4682; S6140; S6163; S668; S7710; S860; S8803a; I719; I534; I517; I526; I518; I863	0.000	5.049	808.49	-268.17	7
653	110219	S10529; S10643a; S1166; S2191; S2259a; S242; S2432; S2510a; S2866; S3356a; S3575a; S4493; S4751; S508; S5427a; S5727a; S5823; S6220; S6345; S6523; S658a; S7031; S7601; S7772a; S7872; S9267; S9421; S9956; I60; I873; I878; I650; I914; I74; I12; I240; I780	10.414	15.560	800.59	31.30	8
476	134712	S10349; S1382; S1982; S2763; S2956; S3221; S3348; S3590a; S3717; S3859; S4541a; S5270; S7300; S7786; S8851a; I64	5.045	7.621	765.27	299.22	9
192	110608	S1956a; S2092; S2377; S2526; S2545; S2556a; S3644a; S3683a; S3721a; S3754; S3839a; S3930a; S47a; S605; S6491a; S6510a; S7310; S7411a; I888; I850; I276; I160; I797; I209	9.980	14.698	709.73	240.78	10

Figure 3.7 Dashboard Illustrating Top 10 Fixed Corridor Results.

3.3.5 Results from Corridor Search

For Corridor Search, a 5-mi corridor length and 0.1-mi increment lengths to move the corridors were provided as input, along with other input parameters discussed in Section 3.3.2. The results were stored in the respective View tables, and analytics were presented using a dashboard. The results of corridor search are presented in table and map format. Similar to the peak searching method, drill-down analysis can be performed on corridors to diagnose the crash patterns on the corridor. Unlike screening by using peak searching, sliding window, and fixed corridor algorithms, the geometry creations for Esri maps when using corridor search algorithm required Dynamic Segmentation (Cadkin, 2002) A PL/SQL script was developed to segment the geometry dynamically based on the results of corridor search. The script joined sections of roadway segments, based on routes. From the results of corridor search, corridors geometry were created based on information regarding the route as well as the begin mile and end mile of corridors. For purposes of illustration, results for corridor search screening on a dashboard are shown in Figure 3.8.

Similar to fixed corridor results, the results of the corridor search were compared to results computed using observed crash frequency and crash rate methods. To the best of the authors' knowledge, the current literature does not provide methods for corridor search; in addition, Safety Analyst does not have capability to perform corridor search as part of network screening. To be able to access the results, observed crash frequency and rates were computed for the same corridors used in this study. Table 3.2 shows the top 15 ranked corridors obtained in this study for the proposed EB-adjusted expected crash frequency and the corresponding corridor ranks, using the observed crash frequency and crash rate methods. As before, the ranks obtained

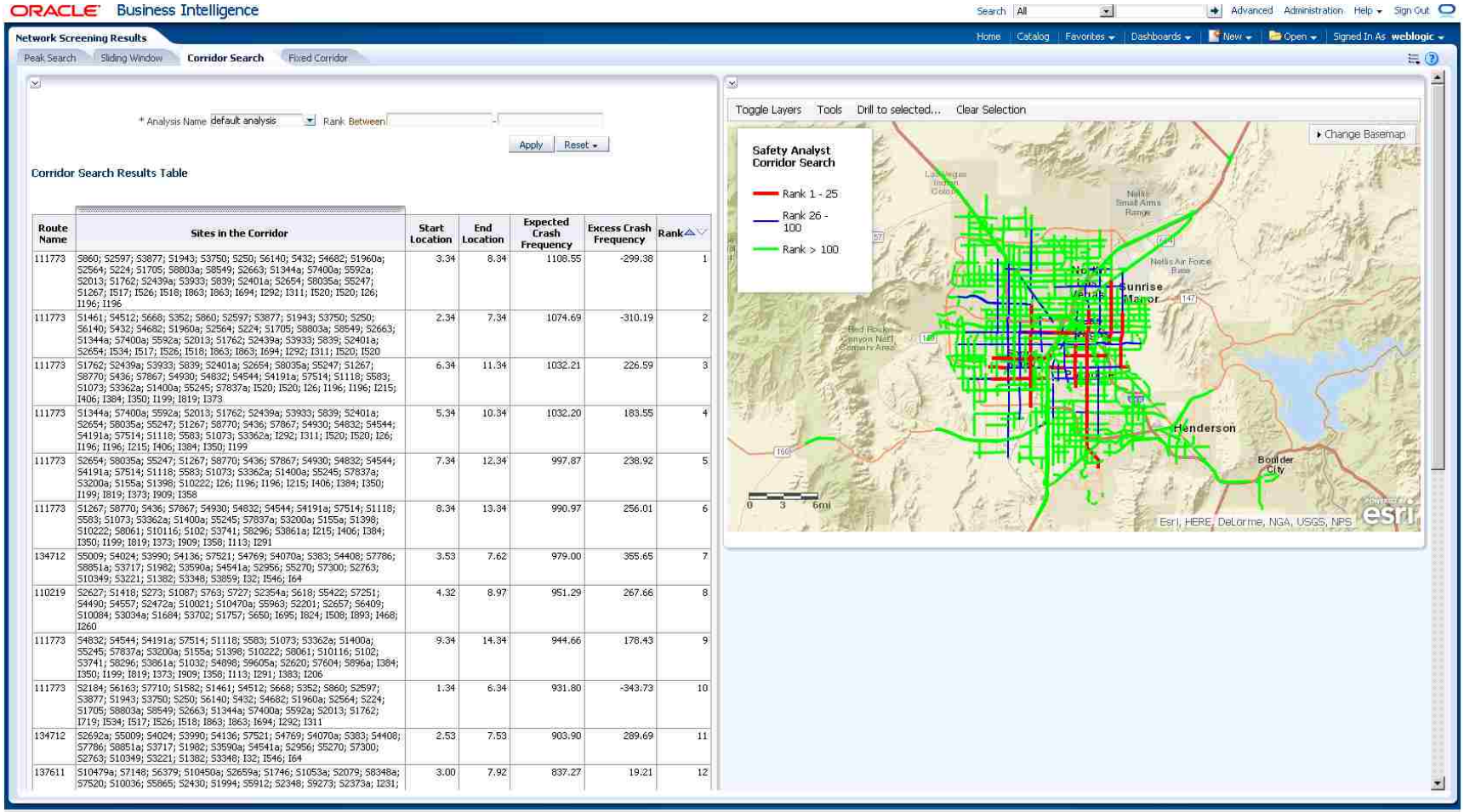
using all the three methods were different due to the use of safety performance functions and the advantages of EB-adjusted expected crash frequency.

Table 3.2 Comparison of Ranks of Top 15 Corridor Search using EB Expected Crash Frequency, Observed Crash Frequency and Crash Rate Methods

Corridor Route	Corridor Begin Mile	Corridor End Mile	Rank - EB Expected Crash Frequency	Rank - Observed Crash Frequency	Rank - Crash Rate
111773	3.339	8.339	1	9	18
111773	2.339	7.339	2	10	19
111773	6.339	11.339	3	14	13
111773	5.339	10.339	4	17	12
111773	7.339	12.339	5	18	17
111773	8.339	13.339	6	19	16
134712	3.525	7.621	7	4	8
110219	4.322	8.974	8	11	23
111773	9.339	14.339	9	19	20
111773	1.339	6.339	10	13	21
134712	2.525	7.525	11	8	11
137611	3.000	7.924	12	5	4
110219	9.213	13.364	13	12	3
110608	8.976	13.016	14	20	9
122385	7.715	11.627	15	3	6

3.4 Conclusions

This research aimed to develop a framework to enable practitioners to use theoretically sound methodologies for network screening for traffic safety analysis. From the perspective of traffic safety engineers, network screening is of significant importance to meet the requirements of a HSIP. Traditionally, separate tools are used 1) to integrate, process, and manage the data; 2) for modeling analysis; and 3) to visualize the results. This traditional approach may result in data replication, and it requires substantial technical knowledge as well as being time consuming. Hence, analysts choose easy-to-implement legacy methodologies, which may lead to identifying incorrectly those sites with safety needs, thus resulting in inefficient roadway-safety management.



Route Name	Sites in the Corridor	Start Location	End Location	Expected Crash Frequency	Excess Crash Frequency	Rank
111773	S860; S2597; S3877; S1943; S3750; S250; S6140; S432; S4682; S1960a; S2564; S224; S1705; S8803a; S8549; S2663; S1344a; S7400a; S592a; S2013; S1762; S2439a; S3933; S839; S2401a; S2654; S1534; I517; I526; I518; I863; I863; I694; I292; I311; I520; I520; I26; I196; I196	3.34	8.34	1108.55	-299.38	1
111773	S1461; S4512; S668; S352; S860; S2597; S3877; S1943; S3750; S250; S6140; S432; S4682; S1960a; S2564; S224; S1705; S8803a; S8549; S2663; S1344a; S7400a; S592a; S2013; S1762; S2439a; S3933; S839; S2401a; S2654; S534; I517; I526; I518; I863; I863; I694; I292; I311; I520; I520	2.34	7.34	1074.69	-310.19	2
111773	S1762; S2439a; S3933; S839; S2401a; S2654; S8035a; S5247; S1267; S8770; S436; S7867; S4930; S4832; S4544; S4191a; S7514; S1118; S583; S1073; S3362a; S1400a; S5245; S7837a; I520; I520; I26; I196; I196; I215; I406; I384; I350; I199; I819; I373	6.34	11.34	1032.21	226.59	3
111773	S1344a; S7400a; S592a; S2013; S1762; S2439a; S3933; S839; S2401a; S2654; S8035a; S5247; S1267; S8770; S436; S7867; S4930; S4832; S4544; S4191a; S7514; S1118; S583; S1073; S3362a; S1400a; S5245; S7837a; I520; I520; I26; I196; I196; I215; I406; I384; I350; I199	5.34	10.34	1032.20	183.55	4
111773	S2654; S8035a; S5247; S1267; S8770; S436; S7867; S4930; S4832; S4544; S4191a; S7514; S1118; S583; S1073; S3362a; S1400a; S5245; S7837a; S3200a; S155a; S1398; S10222; S8061; S10116; S102; S3741; S8296; S3861a; I215; I406; I384; I350; I199; I819; I373; I909; I358	7.34	12.34	997.87	238.92	5
111773	S1267; S8770; S436; S7867; S4930; S4832; S4544; S4191a; S7514; S1118; S583; S1073; S3362a; S1400a; S5245; S7837a; S3200a; S155a; S1398; S10222; S8061; S10116; S102; S3741; S8296; S3861a; I215; I406; I384; I350; I199; I819; I373; I909; I358; I113; I291	8.34	13.34	990.97	256.01	6
134712	S5009; S4024; S3990; S4136; S7521; S4769; S4070a; S383; S4408; S7786; S8851a; S3717; S1982; S3590a; S4541a; S2956; S5270; S7300; S2763; S10349; S3221; S1382; S3348; S3859; I32; I546; I64	3.53	7.62	979.00	355.65	7
110219	S2627; S1418; S273; S1087; S763; S727; S2354a; S618; S5422; S7251; S4490; S4557; S2472a; S10021; S10470a; S5963; S2201; S2657; S6409; S10084; S3034a; S1684; S3702; S1757; S650; I695; I824; I508; I893; I468; I260	4.32	8.97	951.29	267.66	8
111773	S4832; S4544; S4191a; S7514; S1118; S583; S1073; S3362a; S1400a; S5245; S7837a; S3200a; S155a; S1398; S10222; S8061; S10116; S102; S3741; S8296; S3861a; S1032; S4898; S9605a; S2620; S7604; S896a; I384; I350; I199; I819; I373; I909; I358; I113; I291; I383; I206	9.34	14.34	944.66	178.43	9
111773	S2184; S6163; S7710; S1582; S1461; S4512; S668; S352; S860; S2597; S3877; S1943; S3750; S250; S6140; S432; S4682; S1960a; S2564; S224; S1705; S8803a; S8549; S2663; S1344a; S7400a; S592a; S2013; S1762; I719; I534; I517; I526; I518; I863; I863; I694; I292; I311	1.34	6.34	931.80	-343.73	10
134712	S2692a; S5009; S4024; S3990; S4136; S7521; S4769; S4070a; S383; S4408; S7786; S8851a; S3717; S1982; S3590a; S4541a; S2956; S5270; S7300; S2763; S10349; S3221; S1382; S3348; I32; I546; I64	2.53	7.53	903.90	289.69	11
137611	S10479a; S7148; S6379; S10450a; S2659a; S1746; S1053a; S2079; S8348a; S7520; S10036; S5865; S2430; S1994; S5912; S2348; S9273; S2373a; I231;	3.00	7.92	837.27	19.21	12

Figure 3.8 Dashboard Illustrating Corridor Search Results.

In this research, a BI framework is proposed to address barriers associated with data integration, management, and visualization for the implementation of theoretically sound methodologies similar to those in the *HSM*. The outcome is a single framework that accesses the data from the source, integrates and manages the data, processes analytical models, and provides the results by means of a web-based interface. To illustrate the advantages of the proposed framework, network screening algorithms from the *HSM* were implemented and expanded. Results were presented by using dashboards that included maps, filters, and drill downs. Results of network screening produced by this framework were verified by using Safety Analyst and from outcomes by Paz et al. (2015c).

Corridor-level Network Screening was implemented using Fixed Corridor and Corridor Search algorithms. Expected crash frequencies were used instead of observed crash frequencies or rates in order to address regression-to-the-mean bias for selecting corridors with the potential for safety improvements. Top-ranked corridors obtained using the proposed methodology for corridor-level network screening were compared with ranked corridors using rate and frequency methods. The order of ranks of the corridors are completely different as a consequence of using a theoretically sound approach. The advantages of using the proposed framework include the following benefits.

1) It has the capability to perform corridor-level network screening, using a theoretically sound approach.
2) It provides data integration, analysis, and visualization.
3) When new data is loaded into the source, it is automatically loaded into the warehouse, using an ELT process.
4) It has better visualization capabilities than existing methods (Tarko et al., 2014; AASHTO, 2014; Paz et al., 2014).

5) Development cost and time are minimized.

6) Required training and maintenance are minimized.

7) It uses a web-based approach for development and use.

Future work includes automation of dynamic geometry generation for Esri maps for corridor-level network screening. The other three steps of a roadway safety management process also need to be incorporated, which are diagnosis and countermeasure selection, economic analysis and priority ranking, and countermeasure evaluation. Desirable additional capabilities within the proposed framework include methods and tools to 1) estimate SPFs using local data; 2) analyze for diagnosis, countermeasure selection, economic analysis and priority ranking, and countermeasure evaluation to complete safety management process; and 3) perform regional-level forecasting of crash trends. The proposed framework relies on the availability of SPFs. In addition to the outcomes from a standard safety management process, decision makers are required to provide system-wide forecasts and associated targets for long-term planning.

CHAPTER 4

ESTIMATION OF SAFETY PERFORMANCE FUNCTIONS USING CLUSTERWISE REGRESSION

4.1 Introduction

Network screening for sites with the potential for safety improvements is a critical first step in a roadway safety management (RSM) process (Hauer et al., 2002; Montella, 2010; AASHTO, 2010). Network screening could be performed either using 1) traditional methods, such as crash frequencies, rates, and proportions; 2) the state-of-the-art empirical Bayes (EB) method; or 3) the continuous risk profiling (CRP) method. Traditional methods have limitations, including bias associated with traffic volume, segment length, and regression-to-the-mean (Montella, 2010; AASHTO, 2010).

In the EB method, first, a safety performance function (SPF) is used to predict the number of crashes for site types with corresponding traffic volumes and other similar site characteristics. The predicted crash estimates then are combined with the observed crashes to obtain a better estimate of the expected number of crashes. SPFs are crash prediction models that provide estimates of the number of crashes and the associated severity as a function of site characteristics. The EB method addresses the limitations of traditional methods by combining estimates from SPFs, including various site characteristics, and observed crashes. A comparative study by Montella (2010) recommends the EB method as the best for network screening among other existing alternatives.

The CRP method uses a weighted moving average technique (Karlaftis & Tarko, 1998; Depairo et al., 2008) to continuously plot the collision risk profile. The predicted crash

frequency, based on the annual average daily traffic (AADT) for the segment, is obtained from the corresponding SPFs. The SPF value then is transferred to the CRP profile on the same plot. The advantage of the CRP method is that the size of the site is not influenced by the endpoints of the segments. However, this method is not well tested and accepted by researchers.

Appropriate SPFs are essential to determine reliable estimates and avoid bias. In addition, SPFs play a key role in economic analysis as well as the priority-ranking as part of the RSM process (AASHTO, 2010). It is intuitive that a single SPF cannot be used for the entire region or jurisdiction or for all crash types and severities. Similarly, developing SPFs for each possible combination of crash severity, crash type, facility type, and range of explanatory characteristics would require large amounts of data to obtain statistical significant/reliability. It is impractical to take into consideration potential explanatory characteristics because some of them could be expensive to collect, unobservable, or difficult to quantify.

The existing literature classifies sites into several predefined groups as well as site subtypes with measurable and available homogeneous characteristics, such as area type, number of lanes, access control, and median type (e.g., rural two-lane, or urban principal arterial 4-lanes divided.) (AASHTO, 2010). Hence, SPFs for each crash severity and site subtypes are estimated. The assumption is that sites within each subtype experience similar crash patterns as a function of pre-specified explanatory characteristics. That is, the observed pattern of the dependent variable, observed crashes, is not considered explicitly to create site subtypes. For the development of the SPFs, all data points are clustered only as a function of the associated site characteristics.

The consequence could be SPFs with parameters estimated using very different crash patterns. That is, a single function is estimated to represent various distinct trends that could be

captured more accurately by using multiple functions. There could be a number of scenarios; for example, crashes trends for low-volume roads may be different than those for high-volume roads. Similarly, crash trends for low-volume high-speed-limit roads could be different than those for high-volume low-speed-limit roads. Hence, there is potential to create clusters of sites within each predefined site subtype, based on the observed crash trends so as to develop superior SPFs compared to those developed using all data together as a single site subtype. This may lead to an optimal number of SPFs, thus further classifying site subtypes into various sub-groups (clusters) to provide better crash estimates that minimize the overall estimation error.

With advances in data warehousing technology, multiple years of crash data along with associated traffic and roadway characteristics are readily available. The yearly crash and AADT data along with other roadway characteristics constitute panel count data. Karlaftis and Tarko (1998) used a clustering technique on panel crash datasets to account for heterogeneity. Clusters were developed to identify the homogeneous data, and then separate Negative Binomial models were applied to each cluster. Separate models of each cluster provided better results than a joint model. Data was segmented for each of the clusters-based analyses conducted by the authors. However, this technique may not guarantee that each cluster consists of homogeneous sites in terms of crash trends. Many previous studies used clustering analysis to segment the crash data into various clusters so that it could reveal hidden variables that influence crash severity (Depaire et al., 2008; Mohamed et al., 2013; Sasidharan et al., 2015). These studies used latent-class cluster analysis to identify clusters, and then used various types of logit models for modeling crash severity outcomes.

A few studies have estimated the frequency of total crashes using count regression models. The proportion of observed crashes was used to estimate the severity of the crashes,

crash types, light conditions, or vehicles involved in crashes (Geedipally &Lord, 2010; Milton et al., 2008). Wang (2011) used two-stage regression models, first to estimate the crash frequency and then to estimate the crash severity. The advantage of this approach is that traffic and road characteristics data first are used to identify crash frequency based on a full Bayesian approach. Then, more detailed data for individual crashes are used to find the proportion of crashes at various severity levels based on a mixed-logit model. It is clear that clustering was performed for SPFs using either 1) clustering analysis or 2) regression analysis, or 3) stage-wise models, first to perform clustering; then, regression models of crash frequencies were estimated for each cluster.

Most often, segmentation of crash data and classification of SPFs are based on expert knowledge, modeling needs, or the desire to study a specific problem (AASHTO, 2010, Depaire, 2008; Srinivasan, 2013). The selection of a group of sites affects the estimation and reliability of an SPF (Hauer, 2015). Clusterwise regression analysis to simultaneously perform clustering and the generation of the corresponding SPFs so as to minimize the estimation error is lacking in existing traffic-safety literature. Clusterwise regression, introduced by Spath (1979) and extended by Lau et al. (1999), has been used in modeling pavement performance, business, and environment performance (Luo & Chou, 2006; Lu et al., 2014; Poggi & Portier, 2011). This study proposes clusterwise regression to assign sites to clusters and simultaneously seek sets of parameter values for corresponding SPFs so as to maximize the probability of observing the available data. Site membership to clusters and regression parameters are estimated simultaneously to improve the predictive capability of the SPFs. A mathematical programming model is described in detail in Section 4.3 to describe the proposed approach.

4.2 Background of Count Regression Models

Various count data models are available for SPF development and estimation. The most common models are either Poisson or Negative Binomial (NB). If the appropriate data count model is not considered, the statistical validity of the analysis is compromised (Lord & Mannering, 2011). Recent literature has described other models, including Poisson-lognormal, Zero-inflated, Conway-Maxwell-Poisson, Negative Multinomial and Random Parameters (Lord & Mannering, 2011).

Poisson regression models are estimated by specifying the expected number of crashes per time period as a function of explanatory variables. The Poisson model assumes an equal mean and the variance of the crash counts, which often is not correct. In such cases where variance exceeds the mean, referred to as overdispersion, NB models are expected to provide better parameter estimates. The NB model assumes that the expected number of crashes per period follows a gamma probability distribution, and an associated error term is gamma-distributed. In Poisson-lognormal models, the error term is assumed as lognormal. Poisson-lognormal models often are influenced by small sample sizes, and do not have a closed form solution (Miaou et al., 2003). Zero-inflated Poisson or NB models are used when the data is characterized by a substantial number of zeros. Segments having zero numbers of observed crashes are assumed to have long-term mean equal to zero, which is not correct; that is, there may be a crash in the process generating crash data over the long term (Lord & Mannering, 2011).

Conway-Maxwell-Poisson model was proposed to handle over- and under dispersion (mean exceeds variance) characteristics of the crash data. A comparative study by Lord et al. (2008) found that the results obtained by using NB and Conway-Maxwell-Poisson models were

similar. Although Conway-Maxwell-Poisson models can handle under-dispersion data, the results are influenced by low sample means and small sample bias (Lord & Mannering, 2011). In fixed-parameter models, parameter estimates are fixed across observations. If some parameters vary across observations, then a random parameter model can be used to estimate parameters that vary across observations, based on a prespecified distribution. These models provide the better statistical fit than above mentioned models; however, a more complex model estimation is required. A few previous studies have illustrated that this type of model may not improve predictive capability (Lord & Mannering, 2011; Shugan, 2006; Washington et al., 2010).

When data is associated with multiple time periods, the regression models mentioned above cannot be used. For example, crash and AADT data available for five years are aggregated over time. The SPFs developed using aggregated data could underestimate the overdispersion parameter (Kweon & Lim, 2012). With the use of Negative Multinomial (NM) model, yearly crash data and AADT data can be used as multiple time period data (panel data). This prevents the loss of information by not smoothing the AADT over a certain time period (Hauer, 2015). Ulfarsson et al (2003) compared NM models with NB models for predictive modeling and reported that NM models converge with significant higher log-likelihood, provides better fit in terms of log-likelihood ratio and outperforms NB models. In terms of overdispersion, NB models provide better behavior than NM models. The reason could be some of the overdispersion may be captured by the NM's model's temporal seriel correlation.

For further details and applications about these models in SPF development, readers are referred to the study by Lord and Mannering (2010). Considering the multi-year data being used in this research, the Negative Multinomial model specification is proposed for the estimation of SPFs.

Ulfarsson et al, and Hauer (Ulfarsson & Shankar, 2003; Hauer, 2004; Hauer, 2015) explained in detail the theory about NM models, and suggested a log-likelihood distribution function, as shown in Equation 4.1 (Hauer, 2015). Consider a panel data with *i* sites and *m* time periods with Poisson distributed observed crash counts *N*. Assuming the distribution as Gamma, extension of the NB can be applied to panel data, which is the NM distribution. The contribution of site *i* to the log-likelihood is:

$$\ln[L_i^*(\alpha, \beta_1, \dots, b)] = bl_i \ln(bl_i) + \sum_{t=1}^{m_i} N_{i,t} \ln\left(\hat{E}\{\mu_{i,t}\}\right) + ln\Gamma\left(\sum_{t=1}^{m_i} N_{i,t} + b\mathrm{l}_i\right) - ln\Gamma(b\mathrm{l}_i) - \left(\sum_{t=1}^{m_i} N_{i,t} + bl_i\right)\ln[(\sum_{t=1}^{m_i} \hat{E}\{\mu_{i,t}\}) + b\mathrm{l}_i] \quad (4.1)$$

where,

i = number of sites

m = number of time periods starting *t* =1

N = observed crash counts for site *i* with *m* time periods

μ = mean of Poisson distributed crash counts

$\hat{E}\{\mu_{i,t}\}$ = predicted number of crashes at site *i* for time period *t*

l = site length

b = shape parameter

4.3 Methodology

4.3.1 Mathematical Program – Problem Formulation

Notation and Definitions:

The following notations and definitions are used in describing the proposed problem formulation:

i Subscript for a site

I Set of roadway sites (segments, intersections or ramps) to be clustered for a safety performance function, indexed $1 \leq i \leq I$

j Subscript for an explanatory variable

J Set of explanatory variables of a safety performance function, indexed $1 \leq j \leq J$

X An I x J matrix with elements x_{ij}, which are the measurements of explanatory variables for site i $\forall$ $i \in I, j \in J$

k Subscript for a cluster of sites

K Number of safety performance functions of sites, indexed $1 \leq k \leq K$

p_{ik} Membership assignment of a site i to a cluster k

P An I x K binary matrix with elements p_{ik} $\forall$ $i \in$ I, $k \in K$

α_k Intercept for SPF in a cluster k, $\forall$ $k \in K$

β_{jk} Coefficient for explanatory variable j of a SPF in a cluster k , $\forall$ $j \in J$, $k \in K$

N_{ik} Observed number of crashes for site i assigned to cluster k $\forall$ $i \in I$, $k \in K$

$\hat{E}\{\mu_{ik}\}$ Predicted number of crashes of a site i using a SPF in a cluster k $\forall$ $i \in I$, $k \in K$

b_k Shape parameter or inverse over-dispersion parameter of underlying gamma pdf of a SPF in a cluster k, $\forall$ $k \in K$

Objective function

$$\text{Max. } \ln(L^*) = \sum_k \sum_{i|p_{ik}=1} \ln[L^*_{ik}(\alpha_k, \beta_{jk}, \ldots., b_k)] * p_{ik} \quad (4.2)$$

Subject to

Constraints

Log-Likelihood Function

$$\sum_k \sum_{i|p_{ik}=1} f(N_{ik}, \hat{E}\{\mu_{ik}\}, b_k) = \sum_k \sum_{i|p_{ik}=1} \ln[L^*_{ik}(\alpha_k, \beta_{jk}, \ldots., b_k)] \quad (4.3)$$

$$\ln(\hat{E}\{\mu_{ik}\}) = \alpha_k + \sum_{j=1}^{J} \beta_{jk} \ln(x_{ij}) \quad (4.4)$$

Membership constraints

$$p_{ik}=\begin{cases}1, & if\ and\ only\ if\ site\ i\ is\ assigned\ to\ safety\ performance\ function\ k;\\ 0, & Otherwise\end{cases}$$

(4.5)

$$\sum_k p_{ik} = 1\ \forall\ i \in I, k \in K \quad (4.6)$$

$$p_{ik} \geq 0\ \ \forall\ i \in I, k \in K \quad (4.7)$$

For this mathematical program, the sites in the entire data are clustered into *K* safety performance functions by maximizing the log-likelihood of a NM distribution function. The decision variables are the number of SPFs, *K*, the parameters of NM models, $\alpha_k, \beta_{jk}, \ldots., b_k$, and the cluster membership, p_{ik}.

Maximization of log-likelihood is used as the objective function, as shown in Equation 4.2. The objective function finds the set of parameter values that maximizes the probability to observe the available data. The constraint, Equation 4.3 provides the log-likelihood distribution function of the count regression model. The constraint, Equation 4.4, provides the count regression model. In this research, the NM log-likelihood distribution function, as shown in Equation 4.1, is used for the reasons explained in Section 4.2. In order to find the parameter values that maximize the log-likelihood function, the predicted number of crashes are estimated by fitting the distribution function of the model to the data using Equation 4.4. This is performed for the identified number of safety performance functions K. The constraints, Equations 4.5, 4.6, and 4.7, ensure that each site is assigned exactly to one cluster (or safety performance function). Membership p_{ik} takes value of '1' if and only if a site *i* belongs to safety performance function *k*. Otherwise, it takes value '0'.

4.3.2 Solution Algorithm to the Mathematical Program

In order to solve the above mathematical program, a simulated annealing (SA) combined with the maximum likelihood estimation (MLE) algorithm was implemented in R programming language. SA was used for clustering the data to estimate membership of clusters, p_{ik}. For each accepted neighborhood clusters, the MLE was employed to estimate the parameters of the safety performance functions, α_k, β_{jk} and b_k. The 'mle2' function available in the statistical software R was used to estimate these parameters (Bolker, 2016). Román-Román et al. (2012) successfully implemented a SA algorithm for the MLE of the parameters of a Gompertz-type process, which assessed the behavior patterns in several fields of application. DeSarbo et al. (1989) applied such an algorithm to solve the clusterwise linear regression problem.

The algorithm developed to solve the clusterwise regression for this study is illustrated in Figure 4.1, and is described as follows:

Step 1. Initialization

Step 1.1 For a given number of clusters (K), randomly assign cluster memberships to sites.

Step 1.2 Set values of initial temperature (T_0), final minimum temperature (T_{min}), cooling rate (ϕ), and the maximum number of neighbors to be generated (N_{max}) at each temperature level. Set iterator to $N = 0$.

Step 1.3 Count the number of observations of all sites assigned to each cluster. If all the clusters have at least the minimum number of sites, then set α_k, β_{jk}, and $b_k = 1$, and go to step 2; otherwise, reassign the cluster memberships until all clusters have at least the minimum number of sites. Let C_N be the valid initial clusters. Set α_k, β_{jk} and $b_k = 1$.

Step 2. Objective function evaluation and initial parameters estimation

Step 2.1 Estimate the predicted number of crashes with default α_k and β_{jk}

Step 2.2 Estimate the log-likelihood function using the observed number of crashes, predicted number of crashes, and parameter b_k.

Step 2.3 Evaluate the objective function; maximize $\ln\left[L_i^*(\alpha_k, \beta_{jk}, \dots, b_k)\right]$ using MLEm and set this value as MLE_N

Step 2.4 For C_N, obtain α_k, $\beta_{jk,}$ and b_k for all K clusters from MLE.

Step 3. Set of neighborhood clusters generation

Create a set of neighborhood clusters randomly near to the previous cluster, using the following steps:

Step 3.1 Randomly select a prespecified number of sites to change their memberships.

Step 3.2 For each of the site selected, assign a new membership by generating a random number $r \sim R(1, K)$. If the new membership is the same as the previous one, regenerate a random number $r' \sim R(1, K)$ until it is different. Repeat this process until the memberships of all selected sites are different than previously assigned.

Step 3.3 Count the total number of sites assigned to each cluster.

Step 3.4 If all clusters have at least the minimum number of sites, go to Step 5; otherwise, repeat steps 3.1., 3.2., and 3.3. until all clusters have at least the minimum number of sites. Let C_{N+1} be the new set of valid neighborhood clusters.

Step 4. Solution search

Step 4.1 Estimate the predicted number of crashes with default α_k, and β_{jk}

Step 4.2 Estimate the log-likelihood function using the observed number of crashes, predicted number of crashes, and parameter b_k.

Step 4.3 Evaluate the objective function; maximize $\ln\left[L_i^*(\alpha_k, \beta_{jk}, \ldots, b_k)\right]$ using MLE, and set this value as MLE_N

Step 4.4 For C_{N+1}, obtain new α_k and β_{jk} for all K clusters from MLE.

Step 4.5 Calculate $\Delta MLE = MLE_{N+1} - MLE_N$

Step 4.6 Check the following two conditions:

a. If $\Delta MLE > 0$, accept the current solution, C_{N+1} and the corresponding α_k, β_{jk}, and b_k ; go to Step 5; otherwise, go to Step 4.6b.

b. Generate a random number $r" \sim R(0,1)$. Calculate the acceptance probability, $p_{accept} = exp\left(\frac{-\Delta MLE}{B*T}\right)$, where B is a Boltzmann's constant. If $r" > p_{accept}$, accept the current solution, C_{N+1}, and corresponding α_k, β_{jk}, and b_k ; go to Step 6; otherwise, return to Step 3 to generate a set of new neighborhood clusters.

Step 5. Stopping Criteria

Step 5.1 Repeat Steps 3 and 4 for N_{max} times.

Step 5.2 If $T < T_{min}$, stop the algorithm. Otherwise, multiplying the current temperature by the prespecified cooling rate, ϕ, set N =1, and go to Step 2.

The SA algorithm seeks an optimum solution using a probabilistic approach for a given function. Annealing corresponds to progressing a material to its equilibrium state, a process that causes the diffusion of atoms by heating followed by cooling; SA works using a similar technique. Initially, at a high temperature, T, the probability of accepting a worse solution is high. This enables the solution to escape from local maxima, moving downhill as it explores the solution search space vertically as well as horizontally with big step lengths. As temperature cools down, T drops; at this stage, the algorithm uses small step lengths to search heuristically

for an optimum solution on the most promising search space. Román-Román et al. (2012) illustrated that the algorithm converges to a global minimum with a substantially slow cooling rate.

4.4 Experiment and Results

4.4.1 Data Resources and Preparation

Development of safety performance functions requires key data, including information about crashes, traffic flow, traffic control, and roadway characteristics. The data used in this study were extracted from the Nevada Citation and Accident Tracking System (NCATS) database, the Highway Performance Monitoring System (HPMS), and Traffic Records and Information Access (TRINA) of the Nevada Department of Transportation (NDOT) (NDOT, 2016). In addition, the Travel Demand Model (TDM) and Intersections database from Regional Transportation Commission of Southern Nevada (RTC-SN) were accessed. The data consisted of roadway, traffic, intersection, and crash characteristics collected in the Clark County, largest region of the State of Nevada.

A comprehensive database was developed by integrating all the data sources listed above. Various issues were encountered during the development of database, including the availability of data, requirement of data from multiple agencies, and the consistency of the collected data. It was a substantial task to identify data, integrate them, and develop the database.

Various tools were used to integrate the data. Some of data were integrated using location reference system, such as county, route, and milepost. For example, a site was represented as a roadway segment in Clark County on the Route 123 from Milepost 1.300 to 2.500. Crashes were mapped onto this site with the same county and route information for mileposts between 1.300

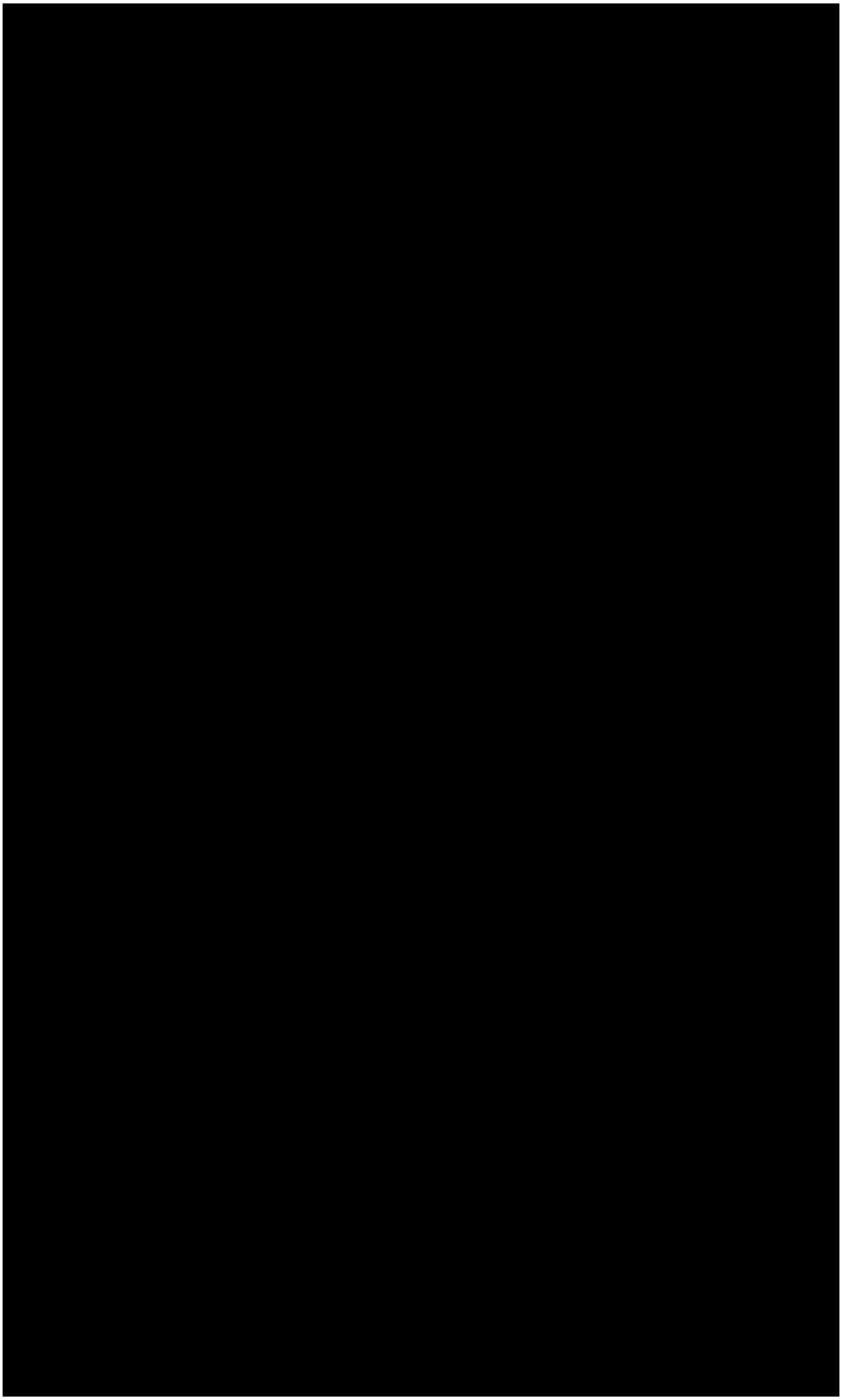


Figure 4.1 Algorithm for Clusterwise Regression to Estimate SPF Parameters.

and 2.500. Other sources where such information was not available were integrated using spatial integration with the help of ArcGIS (ESRI®). For example, crashes mapped onto the intersection specifically required such GIS operations as buffering the signal-control intersection with a 200-ft radius, and mapping crashes within that buffer as intersection-related crashes. The details about the development of this comprehensive database are provided by Paz et al. (2015c).

The database that was developed contains five years of data from 2007 to 2011; it contains 10,287 roadway segments, 973 signal and stop-control intersections, and 973 ramp segments. Few of the consecutive roadway segments indicated homogeneous characteristics. As a result, the data was post-processed to combine the roadway segments with homogeneous characteristics. The characteristics considered were functional class, number of lanes, median type, median width with 5-ft thresholds, speed limits with 5-mph thresholds, and Annual Average Daily Traffic (AADT) with 20% thresholds. The data were classified based on site subtypes that were regularly used in the literature (Hauer 2002; AASHTO, 2010; Hauer, 2015; Paz et al., 2015c).

For this study, site subtypes with a minimum of 70 sites were chosen. Explanatory variables considered in the analysis were segment length, AADT, number of lanes, access control, functional class, median width, median type, posted speed, and terrain type. Variables related with crashes that were considered in this study were the number of crashes and crash severity. Site subtypes identified in the data, based on the literature, are shown in Table 4.1. The post-processing of data and the identification of site subtypes were generated systematically by R code.

Table 4.1 Site Subtypes for Safety Performance Functions

No.	Site Subtype	Description
1	Rural 2-lane	Segments on rural 2 lane road
2	Rural multilane undivided	Segments on rural 4+ lanes with no median
3	Rural freeway 4 lane	Segments on rural 4+ lane with no access control
4	Rural freeway in interchange area	Segments on rural 4+ lane with no access control and in interchange influence area
5	Urban 2-lane arterial	Segments on urban 2 lane arterial
6	Urban multilane undivided	Segments on urban 4+ lanes with no median
7	Urban multilane divided	Segments on urban 4+ lanes with median
8	Urban One-way arterial	Segments on urban one-way arterial
9	Urban freeway 4 lane	Segments on urban 4+ lane with no access control
10	Urban freeway in interchange area	Segments on urban 4+ lane with no access control and in interchange influence area
11	Urban 3-leg signalized	Urban 3-leg Intersections with signal control
12	Urban 4-leg signalized	Urban 4-leg Intersections with signal control
13	Urban 3-leg minor road stop	Urban 3-leg Intersections with stop control on minor roads
14	Urban 4-leg minor road stop	Urban 4-leg Intersections with stop control on minor roads
15	Urban 4-leg all-way stop	Urban 4-leg Intersections with stop control on all roads
16	Urban Diamond off-ramp	Diamond off-ramps in urban area
17	Urban Diamond on-ramp	Diamond on-ramps in urban area

4.4.2. Parameters of the algorithm

Performance of an SA algorithm is affected by the optimization parameters used to solve a specific problem. Selecting an appropriate strategy to generate the neighborhood solution and an appropriate annealing parameters to attain the optimal solution is critical (AASHTO, 2011; Roshan et al., 2013). The literature indicated various methods to determine annealing parameters, including sensitivity analysis (Park et al., 1998; Kirkpatrick et al., 1983; Collins et al., 1988; Rose et al., 1990; Selim & Alsultan 1991; Guo & Zheng, 2005). Previous experience by this research team was drawn up to solve relevant problems in determining the optimization parameters (Paz et al., 2015a; Paz et al., 2015b). In addition, a sensitivity analysis was performed to select the appropriate annealing parameters, taking into consideration a reasonable amount of computation time to reach the optimum solution (Park & Kim 1998). SA parameter values used in this study are shown in Table 4.2.

Table 4.2 Parameters used for optimization in Simulated Annealing

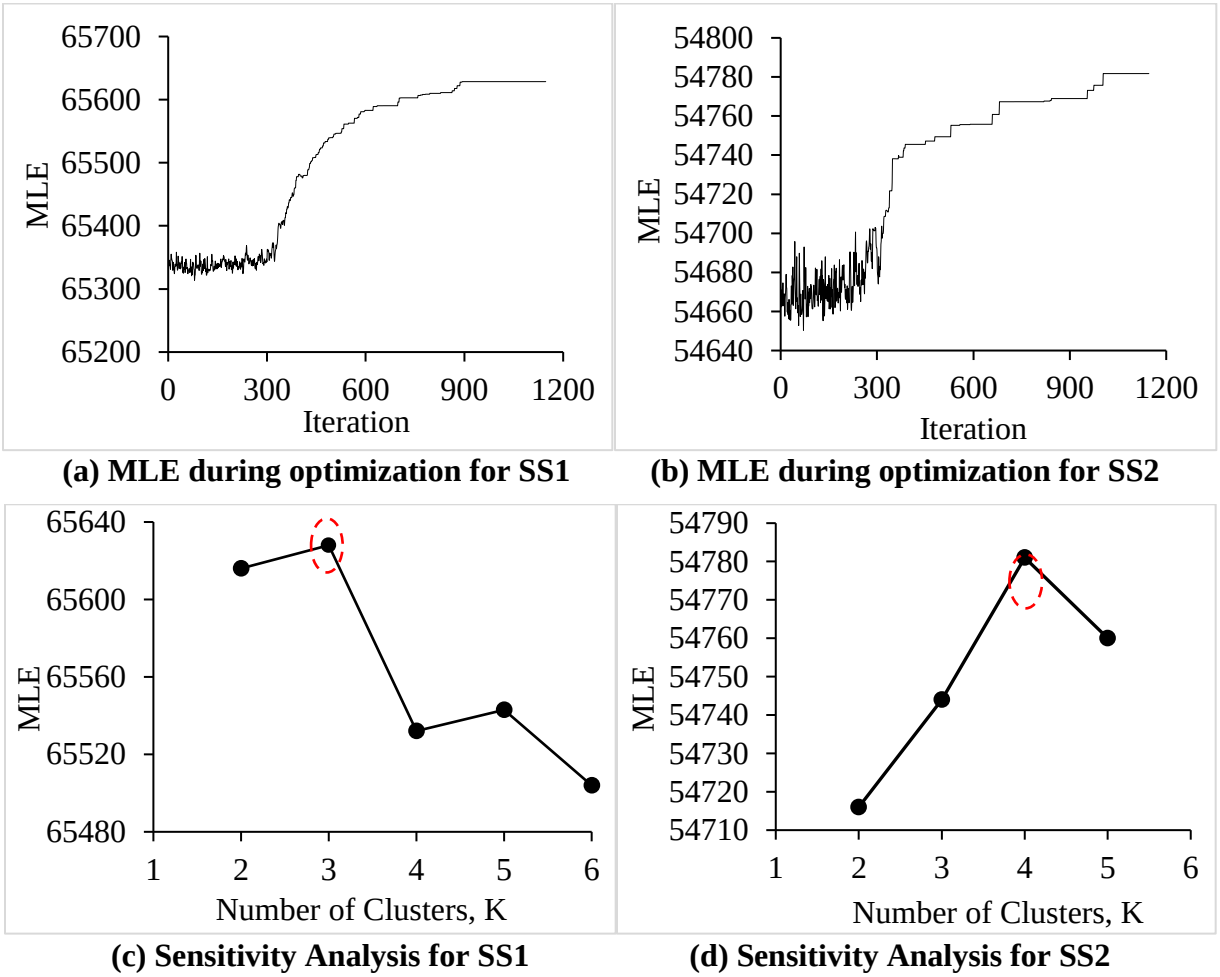
Parameter	Value	Description
T_0	10	Initial temperature
T_{min}	0.0001	Minimum temperature
$_B$	4	Boltzmann constant
Φ	0.99	Cooling rate
N_{max}	10	Maximum number of neighborhood solutions to be generated at each temperature level

4.4.3 Results and Discussion

The results of the two site subtypes were analyzed, urban multilane divided arterial segments (SS1) and urban 4-leg signalized intersections (SS2). These site subtypes were classified into various subgroups (clusters), using clusterwise regression to determine the SPF parameters (total number of crashes) that could provide better crash estimates. Preliminary investigation of crash data – along with explanatory variables of AADT, speed limit, and median type – determined the three clusters that could provide the optimal number of clusters hypothesized for urban multilane divided arterial, SS1. For urban 4-leg signalized intersections, SS2, based on AADT, median type, and speed limit, four clusters could be the hypothesized as the optimal number of clusters. However, the number of clusters hypothesized were confirmed based on the sensitivity analysis.

The algorithm partitioned the data and provided the memberships for sites in these clusters. Figures 4.2a and b indicate the trajectory of the objective function (MLE) when the clusterwise regression models were used for SS1 and SS2, respectively. Figures 4.2c and d show results from the sensitivity analyses for determining the optimum number of clusters for SS1 and SS2, respectively. For SS1, the initial value of maximum likelihood was 65329. After 1,146 iterations, the final value increased to 65628. In the case of SS2, the initial value of maximum likelihood was 54672; the final value increased to 54781.

The optimal number of clusters were further verified using the Bayesian information criteria (BIC), which penalizes the inclusion of additional parameters. Results with BIC values close to negative infinity are categorized as optimal in terms of the number of clusters (Schwarz, 1978; R-Language, 2011). Table 4.3 shows the number of clusters and the corresponding BIC values. The lowest BIC values determine the same optimum number of clusters as identified using sensitivity analysis in Figure 4.2(c) and (d) for SS1 and SS2, respectively.



(a) MLE during optimization for SS1 **(b) MLE during optimization for SS2**

(c) Sensitivity Analysis for SS1 **(d) Sensitivity Analysis for SS2**

Figure 4.2 (a and b) Evolution of MLE during Optimization for a Clusterwise Regression Model and (c and d) Sensitivity Analyses for the Number of Clusters.

In both SS1 and SS2, there is substantial difference in coefficients of parameters of the sites assigned to clusters. In the case of SS1, the coefficients of AADT across clusters are significantly different in magnitude. In the case of SS2, Major Rd AADT and Minor Rd AADT show difference in magnitude across clusters. In addition, other parameters also contribute in clustering the sites with minor difference in magnitude across clusters. The significance level for the parameters was set as 5%. From Table 4.4, it can be observed that not all the parameters were significant, i.e., with p-values less than 0.05. However, for SS1, segment length, AADT, speed limit, and functional class were significant variables in all clusters. A divided median type was not significant in all clusters. This could be due to the collinearity of median width with median type. Functional Class 7 (local road) is not significant, which could due to the very few local road segments in the data. All the parameters in a single cluster model are significant. In the case of SS2, few parameters associated with major road speed and minor road speed had p-values larger than 0.05.

Table 4.3 Results of BIC for Clusters

SS1		SS2	
Clusters	BIC	Clusters	BIC
1	-130542	1	-221951
2	-131046	2	-222309
3	**-131053**	3	-222423
4	-130792	4	**-222573**
5	-130745	5	-222488
6	-130598		

The cluster information and related sites were geo-coded in a map, as shown in Figure 4.3, to investigate any associated geographic pattern among the clusters of sites. However, from Figure 4.3, it was found that there is no specific geographic pattern across the clusters of sites. In

Table 4.4 Estimated Parameters Using the Proposed Clusterwise Regression and the Single-Cluster Method

Urban Multilane Divided Segments (SS1)					Urban 4-leg Signalized Intersections (SS2)					
	Clusters from Clusterwise Regression Method			Single Cluster		Clusters from Clusterwise Regression Method				Single Cluster
Parameters	**Cluster 1** $\eta = 620$	**Cluster 2** $\eta=638$	**Cluster 3** $\eta=633$	$\eta = 1891$	**Parameters**	**Cluster 1** $\eta = 82$	**Cluster 2** $\eta = 86$	**Cluster 3** $\eta = 82$	**Cluster 4** $\eta = 90$	$\eta = 340$
Intercept (Scale)	774.770	853.640	24.400	618.439	Intercept (Scale)	110.950	26.762	6.465	28.907	51.261
length	0.970	0.928	0.824	0.913	Major Rd AADT	0.494	1.247	0.091	0.340	0.538
AADT	1.959	1.219	0.372	1.095	Minor Rd AADT	1.004	0.092	0.152	0.189	0.429
Road Type – State Rd	0.850	0.848	0.565	0.671	Major Rd Lanes = 3-4	1.088	1.253	1.057	1.239	1.225
Functional Class 4	1.092	0.880	0.741	0.891	Minor Rd Lanes > 4	1.252	1.005	2.418	1.571	1.463
Functional Class 6	1.266	0.836	0.543	0.796	Minor Rd Lanes = 3-4	0.709	1.261	1.244	1.589	0.979
Functional Class 7	1.000	0.374‡	0.332‡	0.420	Minor Rd Lanes > 4	0.562	2.059	0.495	2.522	1.020
Number of Lanes 5	0.670	0.449	2.266	1.173	Major Rd Median-Divided	0.477	1.372	0.540	1.829	0.903
Number of Lanes 6	0.774	1.437	1.505	1.173	Minor Rd Median-Divided	1.574	0.741	0.893	0.345	0.818
Number of Lanes 7	0.607	1.247	2.312	1.378	Major Rd Speed 26 to 35	1.187	1.636	0.716‡	0.472	0.803
Number of Lanes 8	0.398	0.119	0.666	0.589	Major Rd Speed 36 to 45	1.249	1.166	1.077‡	0.674	0.810
Median–Flush Paved	2.201	1.510‡	0.817‡	0.972	Major Rd Speed 46 to 55	1.429‡	1.000	0.979‡	0.362‡	1.089
Median–Other Divided	1.000	1.000	0.298‡	0.199	Major Rd Speed > 55	1.000	1.000	1.382‡	1.000	1.062
Median Width 4 – 14 ft	0.938	0.563	0.918	0.775	Minor Rd Speed 26 to 35	1.358	0.924	2.103	0.867	1.172
Median Width > 14 ft	1.219	0.499	0.948	0.864	Minor Rd Speed 36 to 45	0.903	1.163	2.728	0.538	1.163
Speed Limit <= 25	1.218	0.427	2.932	0.898	Minor Rd Speed 46 to 55	1.257‡	0.278	0.788‡	1.201‡	0.922
Speed Limit <= 45	1.239	0.310	2.199	0.750	Minor Rd Speed > 55	1.000	0.656	4.800	1.000	1.256
Speed Limit <= 55	0.609	0.114	2.706	0.503						
Speed Limit > 55	1.490	1.000	1.929	0.869						
Log Likelihood	29675.2	22874.2	13079.1	65304		18367.1	13921.0	10116.5	12377.1	54628
Freeman-Tukey R^2 (R^2_{FT})	0.91	0.85	0.87	0.85		0.93	0.98	0.94	0.93	0.86

addition, histogram of each explanatory variable, for the sites in clusters, was plotted to understand the influence of any specific variable towards the cluster creation. Results indicated no single variable influenced the creation of clusters. However, the coefficients of explanatory variables across clusters in Table 4.4 illustrates combination of explanatory variables were involved in the creation the clusters.

Figure 4.3 Map with color-coded clusters of sites for SS1, Urban Multilane Divided Arterials.

To investigate the performance of the proposed method, the accuracy of the clusterwise models were compared to models developed (single cluster) using the traditional cluster method. All the sites in the SS1 modeled as single cluster were used in the regression analysis to estimate the parameters of the safety performance function. Results obtained from the single cluster are presented in Table 4.4.

The number of sites (samples) in each cluster should be sufficient to obtain statistically reliable estimate of parameters. Larger number of samples in a cluster leads to increased precision during the estimation of parameters using regression. Sample size is estimated using Equation 4.8 (Berenson, 2014). For SS1 and SS2, the minimum number of samples required with 95% confidence interval and 0.1 significance level, are 91 and 75, respectively. The number of sites assigned to each cluster for both SS1 and SS2 are higher than the minimum number of samples required.

$$n = \frac{Z^2 N}{(Z^2 P(1-P)/\varepsilon^2)+(N-1)} \quad (4.8)$$

where,

n = the number of samples,

Z = the Z value for confidence interval,

P = the true parameter, the maximum variance of distribution, 0.5

ε = the significance level,

N = the population size

Model fit for the SPFs were assessed with Goodness-of-Fit (GOF) statistic, Freeman-Tukey R2 coefficient of determination (R2FT) using Equation 4.9. Freeman and Tukey (Freeman, & Tukey, 1950) developed the variance stabilizing transformation for the Poisson distribution as shown in Equations 4.10, 4.11, 4.12 and 4.13. Fridstrøm et al. (1995) applied this transformation when developing generalized regression models for crash data by using Equation 4.9. For the assessment of GOF for models associated with crash data, this statistic is being used by previous studies (AASHTO, 2011; Hamidi, 2010).

$$R_{FT}{}^2 = 1 - \frac{\sum_i^n \hat{e}_i{}^2}{\sum_{i=1}^n (f_i - \bar{f})^2} \quad (4.9)$$

$$f_i = \sqrt{Y_i} + \sqrt{Y_i + 1} \quad (4.10)$$

$$\bar{f} = \frac{\sqrt{Y_i} + \sqrt{Y_i + 1}}{\eta} \quad (4.11)$$

$$\hat{f}_i = \sqrt{4\hat{Y}_i + 1} \quad (4.12)$$

$$\hat{e}_i = f_i - \hat{f} = \sqrt{Y_i} + \sqrt{Y_i + 1} - \sqrt{4\hat{Y}_i + 1} \quad (4.13)$$

where,

Y_i = the observed crashes of site i in a cluster,

$\hat{Y}_i$ = the predicted crashes of site i in a cluster, and

η = the number of sites in a cluster.

Memberships of the sites determined by the algorithm for clusterwise models as well as the associated parameters of the SPF were used to estimate the predicted number of crashes for a clusterwise model. Parameters of the SPF from a single cluster model were used to estimate the predicted number of crashes for a single cluster model. Freeman-Tukey R^2 (R^2FT) are provided

in Table 4.4. For both SS1 and SS2, R^2FT of clusterwise regression SPFs are higher than that of single cluster SPFs.

4.4.3.1 Discussion on Model Overfitting Issue

In addition to the GOF statistic, potential overfitting needs to be investigated. In clusterwise regression modeling, overfitting is a potential issue, which was first highlighted and analyzed by Brusco et al (Brusco, 2008). The study analyzed the effect of the use of explanatory variables to explain variation in the response variable.

This research adopted the approach developed by Brusco et al. (2008) to investigate the presence of overfitting. The total sum of squares (TSS), which is the variation of the response variable about its mean should be equal to between-clusters sum of squares (BCSS) and within-clusters sum of squares (WCSS). The WCSS is equal to the sum of, regression sum of squares (SSR) and sum of squared error of prediction (SSE). The SSR is within-cluster variation explained by regression models and the SSE is the residual error in the clusters. As this study used count data, the TSS, BCSS, SSR and SSE were calculated using Equations 4.10, 4.11 and 4.12. Based on these transformations, the TSS, BCSS, WCSS, SSR and SSE are given by Equations 4.14, 4.15, 4.16, 4.17 and 4.18 respectively.

$$TSS = \sum_{i=1}^{n}(f_i - \overline{f})^2 \quad (4.14)$$

$$BCSS = \sum_{k=1}^{K}\eta_k(\bar{f}_k - \overline{f})^2 \quad (4.15)$$

$$WCSS = \sum_{k=1}^{K}\sum_{i\in k}\left(f_i - \overline{f}_k\right)^2 \quad (4.16)$$

$$SSR = \sum_{k=1}^{K}\sum_{i\in k}\left(\bar{f}_k - \hat{f}_k\right)^2 \quad (4.17)$$

$$SSE = \sum_{k=1}^{K}\sum_{i\in k}\left(f_i - \hat{f}_k\right)^2 \quad (4.18)$$

The TSS, BCSS, WCSS, SSR, and SSE components were calculated for the optimum number of clusters for SS1 and SS2 as shown in Table 4.5. The results illustrated that, for SS1, BCSS is equal to 0.05% of TSS and SSR is equal to 79.19% of WCSS. For SS2, BCSS is equal to 2.11% of TSS and SSR is equal to 68.96% of WCSS. The BCSS accounts only for variation in the response variable by clustering and the SSR accounts for variation in the response variable due to explanatory variables. Obtaining lower percentage for the BCSS and higher percentage for SSR in WCSS indicates that there is no overfitting as most of the variation in the response variable is explained by clustering of response variable with the use of explanatory variables.

Table 4.5 Measures of Overfitting Components Associated with Clusterwise Regression

Measure	Clusterwise Regression - SS1	Clusterwise Regression - SS2
Total sum of squares (TSS)	77,629 (100%)	18,573 (100%)
Between-clusters sum of squares (BCSS)	39 (0.05% of TSS)	392 (2.11% of TSS)
Within-clusters sum of squares (WCSS)	77,590 (99.95% of TSS)	18,181 (97.89% of TSS)
Sum of squares due to regression (SSR)	61,482 (79.19% of WCSS)	12,809 (68.96% of WCSS)
Sum of squared error of prediction (SSE)	16,108 (20.76% of WCSS)	5,372 (28.93% of WCSS)

4.4.3.2 Discussion on prediction accuracy

Predicted crashes then were compared with the observed crashes, and the RMSE for predictions were calculated for both the models (Figures 4.4 and 4.5). The proposed clusterwise regression method was found to perform better than the single cluster method. SPFs in clusters of the clusterwise method had a lower value of RMSE in prediction compared to that of the single cluster method for both SS1 and SS2. In addition, the results showed that using the proposed clusterwise method, the predicted crashes were closer to the 45-degree line, compared with the corresponding prediction using the single cluster method. This indicates that predicted crashes

were closer to observed crashes using the proposed method. This trend is clearly seen for higher numbers of observed crashes.

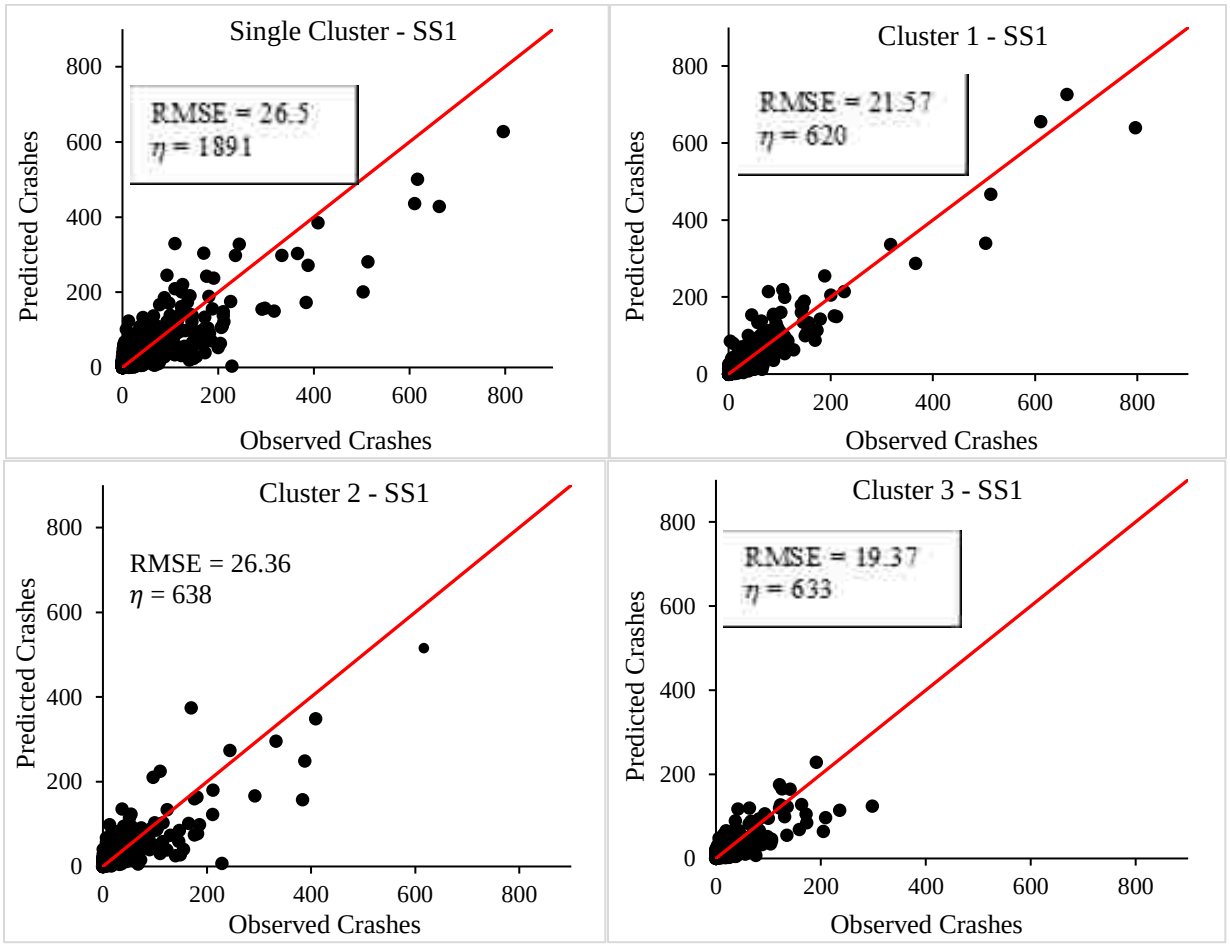


Figure 4.4 Comparison of the Predicted and Observed Number of Crashes, using the Proposed Methods for SS1.

For further validation of the methodology, the dataset is divided into: 1) a test dataset with four years of observations and 2) a validation dataset with one year of observations. Using the proposed methodology, a clusterwise negative multinomial model was developed with the test dataset. Memberships of sites were assigned by mapping sites with memberships determined by clusterwise models. Associated clusterwise models were applied on the validation dataset to estimate the predicted number of crashes. The RMSE values were calculated to compare the prediction accuracy of clusterwise and single cluster models for both the site subtypes, SS1 and

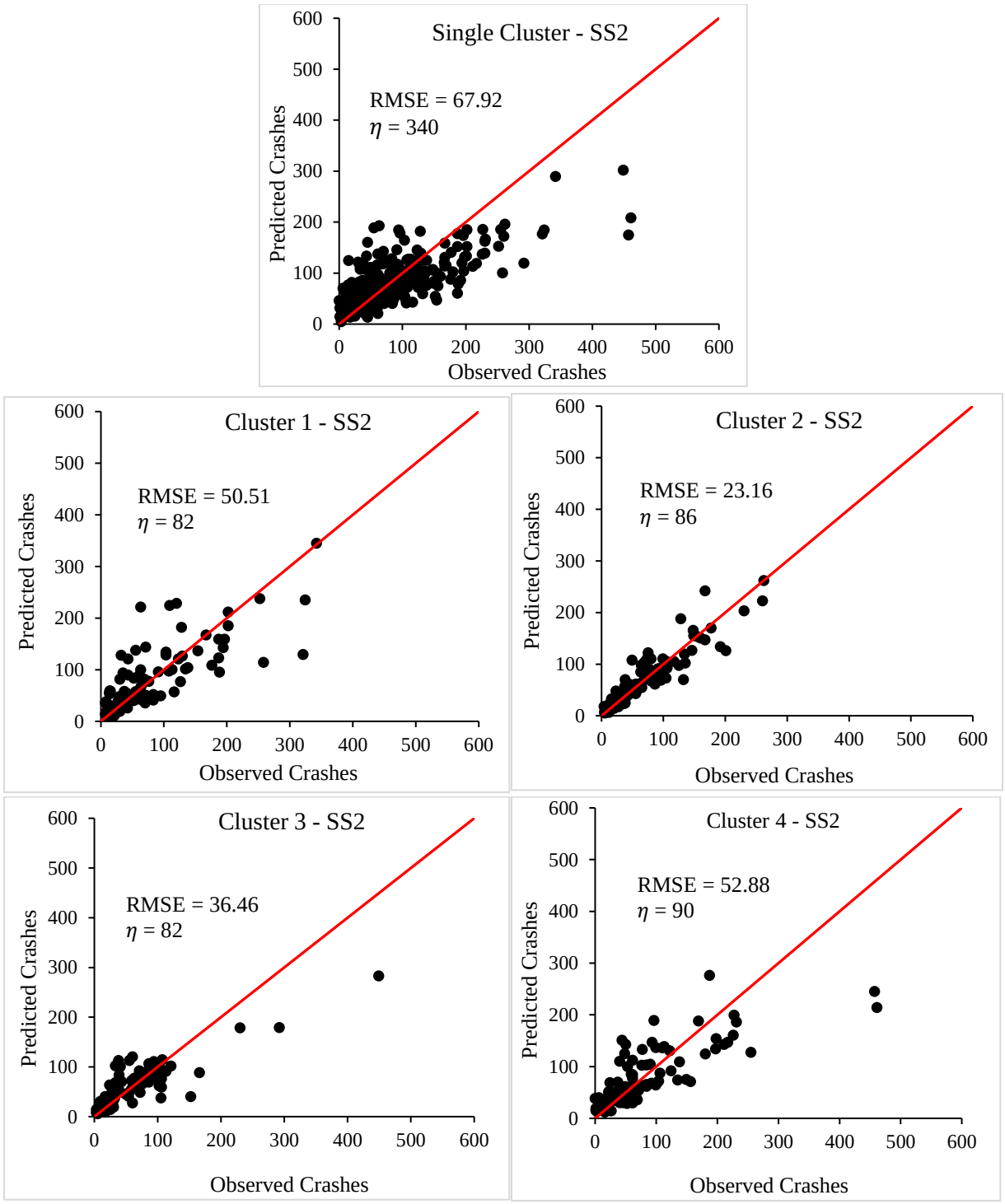


Figure 4.5 Comparison of the Predicted and Observed Number of Crashes, using the Proposed Methods for SS2.

SS2. For SS1, the RMSE values are 5.18 and 5.26 for clusterwise and single cluster models respectively. For SS2, the RMSE values are 10.97 and 11.38 for clusterwise and single cluster models respectively. The lower RMSE values for the clusterwise models indicates the better prediction accuracy in compare to the single cluster model.

4.4.4. Network Screening

The estimated SPFs, using a single cluster model and clusterwise models, were used to estimate predicted crash frequency. The predicted crash frequency of sites were combined with the observed crash frequency to obtain a better estimate of the expected and excess crash frequency using the empirical Bayes method (Montella, 2010). To identify the sites with potential for safety improvements, network screening analyses for arterial roadway segments and intersections were performed using excess crash frequency. It provides a measure of crash frequency at sites where crashes were reduced if a safety improvement was implemented (AASHTO, 2010; AASHTO, 2011). Excess crash frequency was estimated for total crashes, with peak searching on roadway segments having coefficient-of-variation limits for the entire network. With peak searching screening type, a minimum window length of a 0.1 mi segment of the site that had the potential for safety improvement could be determined to deploy a countermeasure. Table 4.6 shows the results of the top 15 sites (the first 15 ranks) having the potential for safety improvements for the estimated SPFs using single cluster and clusterwise regression. It is clearly evident that the sites were ranked in different order when using SPFs from clusterwise regression and SPFs from a single cluster method. In addition, three roadway segment sites (S6570, S8070, and S6043), in Table 4.6, are not ranked by the single cluster method within top 15 sites as was the case with clusterwise regression ranking. Similarly, eight intersection sites (I411, I759, I452, I519, I839, I377, and I580), in Table 4.6, are not ranked by the single cluster method within top 15 sites as

was the case with clusterwise regression. Considering budget constraints, these results could make significant difference for choosing sites with potential for safety improvements.

Table 4.6 Network Screening Results using the Proposed Clusterwise Regression and the Single-Cluster Method for Arterial Roadway Segments and Intersections

Clusterwise Regression				Single Cluster Regression			
Segment/ Intersection ID	Observed Crash Frequency */**	Predicted Crash Frequency */**	Excess Crash Frequency */**	Segment/ Intersection ID	Observed Crash Frequency*/ **	Predicted Crash Frequency*/ **	Excess Crash Frequency*/ **
Roadway Segments*							
S3346	256.00	40.78	213.45	S3346	256.00	35.34	218.57
S9695	270.00	59.50	209.31	S5700a	276.00	56.98	217.73
S4459	240.00	30.43	207.26	S9695	270.00	65.56	203.39
S3521a	234.00	46.12	186.52	S4459	240.00	35.78	202.30
S5700a	276.00	94.59	180.77	S7758	240.00	43.85	194.65
S3129a	204.00	32.96	169.31	S6570	242.00	50.59	190.14
S5658	202.00	35.46	164.97	S3521a	234.00	44.17	188.39
S7736a	210.00	46.15	162.66	S8070	272.13	85.74	185.47
S2374	213.50	32.90	159.91	S7736a	210.00	27.36	180.41
S1775	212.00	56.78	154.30	S5658	202.00	29.10	170.91
S7736a	200.00	46.15	152.73	S7736a	200.00	27.36	170.54
S7693a	192.00	40.83	149.93	S3129a	204.00	32.64	169.60
S2941	172.00	21.77	147.93	S6043	252.00	8.94	167.91
S4843	180.00	32.41	146.07	S7693a	192.00	24.07	165.60
S4648a	176.00	29.00	145.31	S2374	213.50	24.14	160.97
Intersections**							
I900	51.60	23.26	27.60	I703	66.6	24.116	41.41
I905	37.40	11.16	24.34	I411	91.4	52.84	38.11
I703	66.60	45.81	20.51	I759	92.2	55.81	35.99
I139	50.40	30.95	19.07	I452	51.8	23.548	27.25
I65	35.20	16.80	17.74	I900	51.6	28.38	22.72
I551	64.80	48.48	16.12	I135	58.4	36.697	21.34
I634	40.20	23.68	16.09	I248	37.4	18.168	18.59
I135	58.40	42.20	15.96	I86	31.6	14.552	16.35
I643	39.40	23.23	15.75	I519	64.2	47.679	16.31
I779	39.00	23.65	14.96	I839	49.6	33.535	15.77
I816	40.00	25.14	14.50	I377	28.8	12.110	15.57

Clusterwise Regression				Single Cluster Regression			
Segment/ Intersection ID	Observed Crash Frequency */**	Predicted Crash Frequency */**	Excess Crash Frequency */**	Segment/ Intersection ID	Observed Crash Frequency*/ **	Predicted Crash Frequency*/ **	Excess Crash Frequency*/ **
I808	39.20	24.64	14.20	I83	38.4	22.433	15.53
I433	33.80	21.20	12.24	I580	30.8	15.325	14.87
I86	31.60	19.03	12.17	I958	35.4	20.887	14.09
I83	38.40	25.95	12.16	I551	64.8	51.187	13.45

*Segments – units of crash frequency are crashes/mile/year
**Intersections – units of crash frequency are crashes/year

4.5 Conclusions

This study proposed and implemented a clusterwise regression to develop safety performance functions. The objective was to minimize the estimation error by considering multiple SPFs rather than a single SPF for a sample of similar sites. A combinatorial nonlinear mathematical program was formulated. The clusterwise method simultaneously segmented the roadway sites into a number of clusters, and estimated the parameters of the SPF for each cluster. A simulated annealing coupled with maximum likelihood was used to solve the mathematical program. Considering the data characteristics, a Negative Multinomial count model was used for regression, which took into account temporal factors by not aggregating traffic and crashes over the period. The algorithm was tested for number of clusters, and sensitivity analysis was performed to validate that it was an optimum solution for a provided dataset.

The results obtained from the proposed clusterwise models were compared with the results obtained using a single cluster method. The comparison showed that the proposed clusterwise regression method performed slightly better than a single cluster method in predicting crash estimates. With network screening using clusterwise regression, DOTs can use their resources in an efficient manner. The gain in predicting crashes could translate into significant savings in terms of lives and societal costs.

The clusterwise model could be improved by determining the optimum number of clusters as one of the decision variables. The significance of explanatory variables needs to be determined before assigning cluster memberships to roadway sites. This would exclude insignificant variables and reassign the cluster memberships, which may result in an improved estimate of SPFs. Correlation of explanatory variables should be investigated, and some should be removed during the optimization process.

CHAPTER 5

FORECASTING PERFORMANCE MEASURES FOR TRAFFIC SAFETY USING DETERMINISTIC AND STOCHASTIC MODELS

5.1 Introduction

The Federal Highway Administration (FHWA) requires state Departments of Transportation (DOTs) to develop Highway Safety Plans (SHSPs) to obtain funding as part of the two legislative acts, the earlier Safe, Accountable, Flexible, Efficient Transportation Equity Act – A Legacy Users (SAFETEA-LU) and the new act, Moving Ahead Progress in 21st Century (MAP-21) (FHWA, 2015). Performance-based highway safety is one of the ten performance provisions that MAP-21 seeks to use in order to strengthen the U.S. Surface Transportation Program.

The FHWA is seeking criteria to assess traffic safety regarding 1) fatalities per Vehicle Miles Traveled (VMT), 2) serious injuries per VMT, 3) the number of fatalities, and 4) the number of serious injuries. State DOTs and Metropolitan Planning Organizations (MPOs) are required to use these four measures to conduct federal-aid highway programs and assess performance. Each state DOT should set and report targets based on these performance measures for each year, and they should set a goal of achieving these targets within two years. If the targets are not met, or significant progress has not been achieved, DOTs could face funding issues by the obligation authority (FHWA, 2015).

As a primary step to meet these FHWA requirements, state DOTs could focus on three areas: 1) the means to ensure quality data, 2) forecasting capabilities, and 3) methods and tools to automate the management, processing, and easy use of data including the forecasting

capabilities. In addition, DOTs will be required to implement traffic-safety improvement projects based on these three areas.

Regarding acquiring quality data, most DOTs already have committed resources for data collection, processes, and storage. Standards and policies to collect crash data have been based on the U.S. DOT's *Model Minimum Uniform Crash Criteria (MMUCC) Guideline* (MMUCC, 2012). The Nevada Department of Transportation (NDOT) has a crash database, the Nevada Citation and Tracking System (NCATS). In a joint effort by the Office of Traffic Safety and NDOT's Safety Engineering Division, data collected by police officers have been loaded into this database.

For many state DOTs, the development of forecasting methods is in an early stage of research. DOTs' Strategic Highway Safety Plans need to establish statewide performance measures, targets, and strategies to improve traffic safety across critical emphasis areas (Park & Young, 2012). However, few DOTs have developed robust methodologies for forecasting performance measures that help set appropriate targets for the reduction of fatalities and serious injuries (Park & Young, 2012). In various DOTs, traffic safety analysts use extrapolation or simple moving-average models to establish targets (FHWA, 2013).

In many instances, these targets have been set as ideal (aspirational) goals rather than based on approaches using evidence or models (Kweon, 2010). For example, many DOTs have adopted zero-fatality targets to emphasize the importance of traffic safety to the public. The SHSPs of most states have ideal goals, with easily achievable targets set over a five-year period (FHWA, 2013). Considering that crashes are random events, forecasting of traffic-safety performance measures, which are used to set safety goals, should be model-based or evidence-based.

For a traffic safety policy, a model-based approach generally is avoided for forecasting and setting targets. This is because the required data collection is extensive and expensive; in addition, establishing a relationship between the performance measures and influencing factors is difficult (Kweon, 2010). For example, in order to use a model-based approach, the Average Annual Daily Traffic (AADT) is one of the exposure variables that is required for all public roads. Most DOTs only collect the corresponding data for a sample of their facilities.

Although the evidence-based approach also requires data, the demands are fewer compared to the model-based approach. Typically, forecasting models using time-series data are used for the evidence-based approach (Kweon, 2010; Yannis et al., 2011; Zhang et al., 2015; Sukhai et. al, 2011). When quality data is combined with sound forecasting models, DOTs can use the results to deploy strategies that reduce the number of traffic fatalities and injuries. In addition, methods and technologies to automate the management and use of data, such as data warehousing coupled with Business Intelligence, can accelerate and improve the deployment of traffic safety solutions.

Crash information can be characterized as count data. Poisson and negative binomial regression models are used typically to study this type of data. Lord and Mannering (2010) provided a review of several approaches – the generalized estimating equation (GEE), random effects, random parameters, finite mixture, and Markov switching – to model dispersed count data. Given that crash data are repeatedly collected across time, there is a possibility of correlated error terms from adjacent time periods, or serial correlation. Quddus (2008) noted that the count models listed above do not consider the effect of serial correlation found frequently in time-series count data. In order to handle the underlying serial correlation effects of crash count

data, deterministic exponential smoothing models (Brown, 1963; Holt, 1957) and stochastic autoregressive moving average models (Box et al., 2008) can be used.

This study focusses on forecasting models for traffic-safety performance measures by using time-series data from the Nevada Department of Transportation. Specifically, this study proposes a data-driven deterministic and stochastic time-series methodology to forecast traffic-safety performance measures required by MAP-21. The models used in this research are among those commonly applied for time-series crash data. Potentially, other DOTs and MPOs could apply this approach, provided the required data is available.

5.2 Methodology

This study considered four independent and univariate time-series of crash datasets — that were used to forecast several performance measures required by MAP-21: the number of fatalities, the number of serious injuries, fatalities per 100 million VMT, and serious injuries per 100 million VMT. Two datasets were time-series crash counts that were aggregated monthly from 1994 to 2012 (seasonal data). The first one corresponds to the number of fatalities. The second one corresponds to the number of serious injuries. The other two datasets from years 1995 and 2013 aggregated over the exposure variable, VMT, are the annual fatalities per 100 million VMT and serious injuries per 100 million VMT. These datasets are non-seasonal. Because monthly data for statewide VMT was unavailable, seasonal data was not used for fatalities or serious injuries.

An exploratory analysis of all the datasets was conducted to determine the characteristics of each series. To forecast each time series, deterministic and stochastic models were used to examine which was appropriate to use for these datasets. To select the best forecast model, results were compared using as goodness-of-fit the root-mean-square error (RMSE), and the mean absolute

percentage error (MAPE). The model with less MAPE and RMSE values was considered to be optimal.

5.2.1 Deterministic Forecasting

Various models for deterministic exponential smoothing were used in this study, and include 1) Simple Deterministic, 2) Holt, 3) Brown, 4) Damped-trend, 5) Seasonal, 6) Winter-additive, and 7) Winter-multiplicative. The first four models were used in this study to forecast the number of fatalities and the number of serious injuries per 100 million VMT in Nevada for five years from 2013 to 2017. However, because seasonal data was unavailable, the last three models were not used for these forecasts.

The Simple Deterministic model is appropriate for a series that have no trends or seasonality. The Holt and Brown models are appropriate for series with a linear trend and no seasonality. The smoothing parameters for the Holt model are level and trend, which are not constrained; the parameters for the Brown model are level and trend, which are assumed to be equal. The Damped-trend model is appropriate for a series with a linear trend that is dying out without seasonality; its smoothing parameters are level and with a damping trend (Gardner, 1985).

With a Simple Moving Average, observations are weighted equally. Relative to older observations, an Exponential Smoothing model assigns more weight to more recent data. Exponential Smoothing was first suggested by Brown (Brown, 1963) and expanded by Holt (Holt, 1957). Brown's simple exponential smoothing (Brown, 1963) is expressed in Equation (5.1).

Simple Exponential Smoothing does not provide adequate estimates when there is a trend in the data. To handle trends in the data, Double Exponential Smoothing models, such as the Holt and Brown models in Equations (5.2) and (5.3), respectively, have been proposed. Double Exponential Smoothing introduces a term to capture the trend in the forecasting data (Gardner, 1985).

Forecasts by Double Exponential Smoothing models use a constant trend, which may result in over-forecasting for long horizons. A parameter that 'dampens' the trend to a flat line in the future was introduced by Gardner (Gardner, 1985), as illustrated in Equation (5.4).

$$Y_{t+1} = \alpha X_t + (1-\alpha) Y_t \tag{5.1}$$

$$\begin{aligned} Y_t &= \alpha X_t + (1-\alpha)(Y_{t-1} + \theta T_{t-1}) \\ T_t &= \beta (Y_t - Y_{t-1}) + (1-\beta)\theta T_{t-1} \end{aligned} \tag{5.2}$$

$$\begin{aligned} Y_{t+1} &= Y_t + T_t \\ Y_t &= \alpha X_t + (1-\alpha)(Y_{t-1} + T_{t-1}) \\ T_t &= \beta (Y_t - Y_{t-1}) + (1-\beta) T_{t-1} \end{aligned} \tag{5.3}$$

$$\begin{aligned} Y_{t+1} &= Y_t + T_t \\ Y_t &= \alpha X_t + (1-\alpha)(Y_{t-1} * T_{t-1}) \\ T_t &= \beta (Y_t / Y_{t-1}) + (1-\beta) T_{t-1} \end{aligned} \tag{5.4}$$

where:

Y_{t+1} = forecasted value for time period $t+1$

X_t = observed value at time period t

Y_t = forecasted Level value which represents the smoothed value up to time period t

T_t = trend estimate at time period t (slope of the trend line that we are fitting at time period t)

$\alpha, \beta,$= smoothing parameters (should be between 0 and 1)

θ = damping parameter

When a time series contain a seasonal factor, the appropriate models could include a Seasonal, Winter additive, and Winter multiplicative component. The Winter-additive model is illustrated by Equation (5.5).

$$
\begin{aligned}
Y_{t+1} &= (Y_t + T_t) * S_t \\
Y_t &= \alpha \frac{X_t}{S_{t-c}} + (1-\alpha)(Y_{t-1} + T_{t-1}) \\
T_t &= \beta(Y_t - Y_{t-1}) + (1-\beta)T_{t-1} \\
S_t &= \gamma \frac{X_t}{Y_t} + (1-\gamma)S_{t-c}
\end{aligned}
\tag{5.5}
$$

where:

Y_{t+1} = forecasted value for the time period t+1

Y_t = forecasted Level value which represents the smoothed value up to time period *t*

T_t = trend estimate at time period *t (*slope of the trend line that we are fitting at time period *t)*

X_t = observed value at time period *t*

α, β, γ = smoothing parameters

S_t = seasonal parameter estimate

If the series lack a trend, Equation 5.5 without T_t describes the Seasonal model. The Winter-multiplicative model is appropriate if time series has a trend and the smoothing parameters are level, trend, and assumed to be equal.

5.2.2 Stochastic Forecasting

Two stochastic forecasting models were used in this study including the Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models. The ARIMA model uses past values and past errors to capture trends and predict future values. This model was first introduced by Box and Jenkins, and various transformations of this model have been developed (Box et al., 2008). An ARIMA (*p,d,q*) model with a seasonality factor is known as a SARIMA $(p,d,q)(P,D,Q)_s$ model (Nobre et al., 2001).

In this study, an ARIMA model was used to forecast fatalities and serious injuries per million VMT. SARIMA was used to forecast the number of fatalities and serious injuries where seasonality data was available. Multiple combinations were tested for the autoregressive process, moving averages, seasonal factors, and transformations to achieve the stationary condition for each dataset. The general equations of ARIMA (*p, d, q*) and SARIMA (*p, d, q*) $(P, D, Q)_s$ models are presented in Equations (5.6) and (5.7), respectively.

$$\hat{y}_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{5.6}$$

$$\begin{aligned} y_t &= Y_t && \text{for d} = 0 \\ y_t &= Y_t - Y_{t-1} && \text{for d} = 1 \\ y_t &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) && \text{for d} = 2 \end{aligned}$$

The differencing, if any, must be reversed to obtain a forecast for the original series.

$$\begin{aligned} &\text{if d} = 0 \quad \hat{Y}_t = \hat{y}_t \\ &\text{if d} = 1 \quad \hat{Y}_t = \hat{y}_t + Y_{t-1} \\ &\text{if d} = 2 \quad \hat{Y}_t = (\hat{y}_t + Y_{t-1}) + (Y_{t-1} - Y_{t-2}) \end{aligned}$$

where,

Y_t = observed values

y_t = differenced (stationarized series)

$\hat{y}_t$ = Forecast of the stationarized series

$\hat{Y}_t$ = Forecast of the original series

Φ = autoregressive parameters

θ = moving average parameters

Φ_0 = model constant is assumed different from zero

ε_t = error

p = the number of autoregressive terms

d = the number of nonseasonal differences

q = the number of moving-average terms

$$\phi(B)\Phi(B^s)(h_t - \mu) = \theta(B)\Theta(B^s)e_t$$

$$h_t = (1-B)^d(1-B^s)^D Y_t = \Delta^d \Delta_S{}^D (Y_t) \qquad (5.7)$$

where:

B^s = seasonal lag operator $B^s(Y_t) = Y_{t-s}$

Δ^d = $(1\text{-}B)$ difference operator

$\Delta_S{}^D$ = $(1\text{-}B^s)$ seasonal difference operator

h_t = stationary series

Y_t = observed series

B = lag operator

$\phi(B)$ = autoregressive order p (ordinary part of the series)

$\theta(B)$ = moving average order q (ordinary part of the series)

$\Phi(B^s)$ = autoregressive order P (seasonal part of the series)

$\Theta(B^s)$ = moving average order Q (seasonal part of the series)

μ = average of stationary series

e_t = model error

D, d = times they have applied the seasonal difference and regulate difference for the original series

5.3 Results and Discussion

5.3.1 Results for Deterministic Forecasting Models

For the number of fatalities and serious injuries accumulated every month, 228 data points were collected. The mean and standard deviation for fatalities and serious injuries were (27; 200) and (7.98; 41.73), respectively. For fatalities and serious injuries per 100 million VMT data, 19 data points were collected; the mean and standard deviation were (0.42; 3.87) and (1.74; 12.38), respectively. The VMT was obtained from reports prepared by NDOT (Nevada Department of Transportation).

For the number of fatalities and serious injuries, Table 5.1 provides the goodness-of-fit measures for the deterministic models. Considering the MAPE and RMSE, the Winter-additive model was the best deterministic method to forecast the number of fatalities and serious injuries in Nevada. This indicates that the seasonal time-series data had a level and trend, and are constrained with each other.

Table 5.1 Goodness-Of-Fit for the Deterministic Models – Number of fatalities and Serious Injuries

Deterministic Model/ Performance Measures	**Number of Fatalities**		**Number of Serious Injuries**	
	MAPE	RMSE	MAPE	RMSE
Simple	25.698	7.361	12.309	29.891
Holt	25.469	7.374	12.185	29.942
Brown	25.691	7.419	12.681	30.520
Damped Trend	25.703	7.393	12.304	30.025
Simple Seasonal	22.770	6.729	10.724	25.638
Winter Additive	22.503	6.735	10.476	25.593
Winter Multiplicative	23.126	6.890	10.644	26.138

Similarly, from Table 5.2, the Damped Trend was the best method to forecast fatalities and serious injuries per million VMT in Nevada. This indicates that the data had a time series with a linear trend that was dying out.

Table 5.2 Goodness-Of-Fit for the Deterministic Models – Rate of Fatalities and Serious Injuries

Deterministic Model/ Performance Measures	Fatalities/ 100 Million VMT		Serious Injuries/ 100 Million VMT	
	MAPE	RMSE	MAPE	RMSE
Simple	7.714	0.166	8.987	1.262
Holt	6.888	0.158	5.916	0.916
Brown	8.659	0.176	6.300	1.015
Damped Trend	6.867	0.163	5.291	0.911

5.2.2 Results for Stochastic Forecasting Models

The time series for the number of fatalities and serious injuries, had 228 points of data, collected monthly. Multiple combinations of SARIMA models with various *p, q, P,* and *Q* values were applied and tested with seasonal and non-seasonal differenced values. The need for differencing the monthly crash-count data for the number of fatalities and serious injuries were identified by checking stationarity; the autocorrelation function and partial autocorrelation function plots were used to identify *p, d, q, P, D* and *Q* values.

Based on the goodness-of-fit measures, low MAPE values, and low RMSE values, SARIMA $(0,0,5)(0,1,1)_{12}$ was selected as the model with the best fit for forecasting the number of fatalities and serious injuries. Table 5.3 shows the results of the goodness-of-fit for the various SARIMA models for the number of fatalities and serious injuries.

Figure 5.1 and 5.2 illustrates the forecast of number of fatalities and serious injuries, of best fit SARIMA $(0,0,5)(0,1,1)_{12}$ model. Figure 5.3 and 5.4 illustrates the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of this model. The residuals of autocorrelations and partial autocorrelations, near zero, illustrates that it did not significantly differ from a zero-mean. This indicates that the model has good statistical fit with the data (Box, 1970).

Table 5.3 Goodness-Of-Fit for the SARIMA Models

Stochastic Model/ Performance Measures	Number of Fatalities		Number of Serious Injuries	
	MAPE	RMSE	MAPE	RMSE
SARIMA(0,0,5)(0,1,1)$_{12}$	25.120	7.756	12.214	30.111
SARIMA(0,0,4)(0,1,1)$_{12}$	25.910	7.519	12.303	30.113
SARIMA(0,0,3)(0,1,1)$_{12}$	25.665	7.870	12.855	31.489
SARIMA(0,0,2)(0,1,1)$_{12}$	25.662	7.892	13.406	32.663
SARIMA(0,0,1)(0,1,1)$_{12}$	27.017	7.993	13.676	33.132

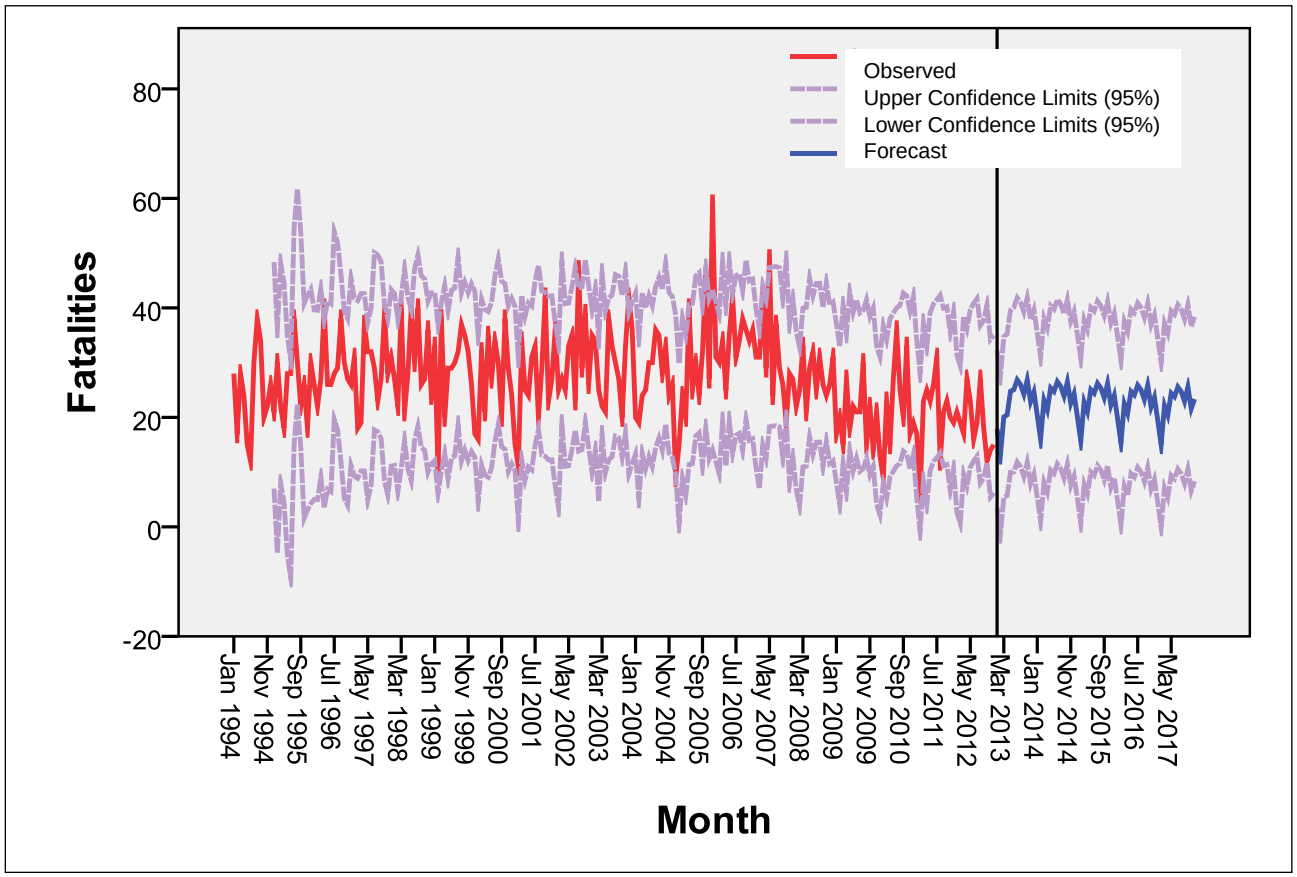


Figure 5.1 Forecast of the Fatalities using the SARIMA(0,0,5)(0,1,1) model.

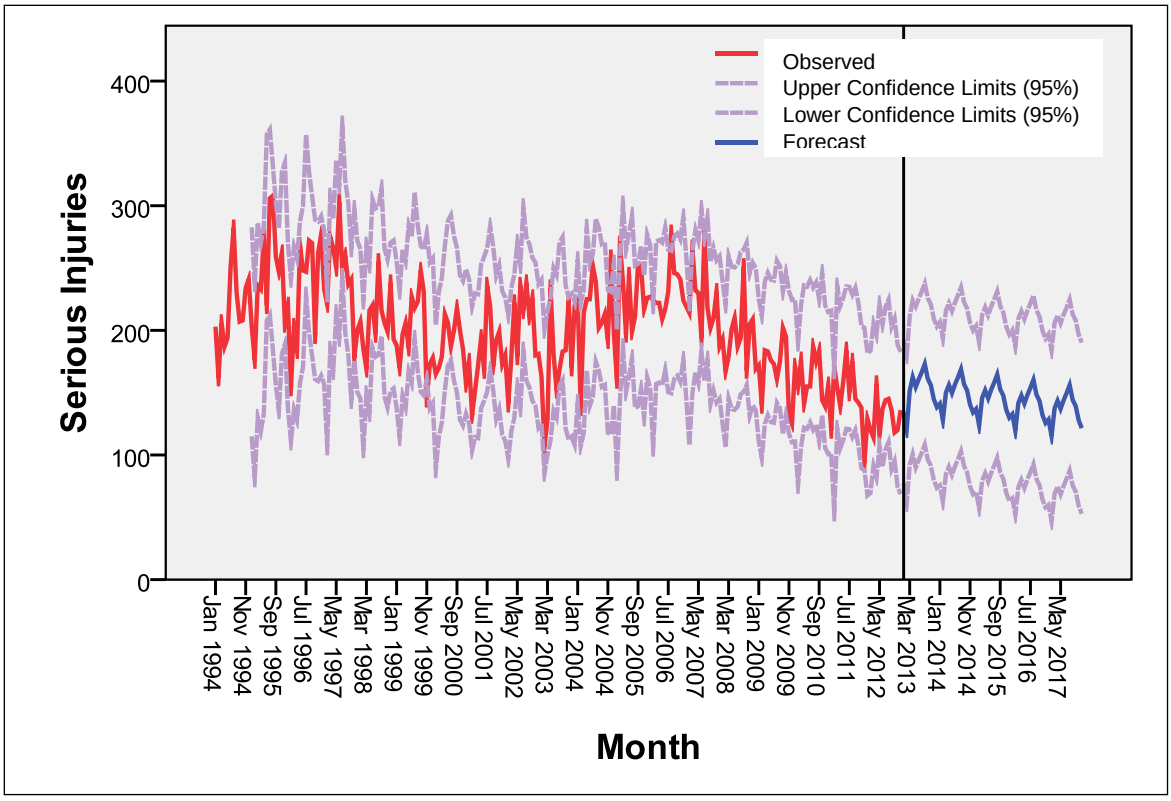


Figure 5.2 Forecast of the Serious Injuries using the SARIMA(0,0,5)(0,1,1) model.

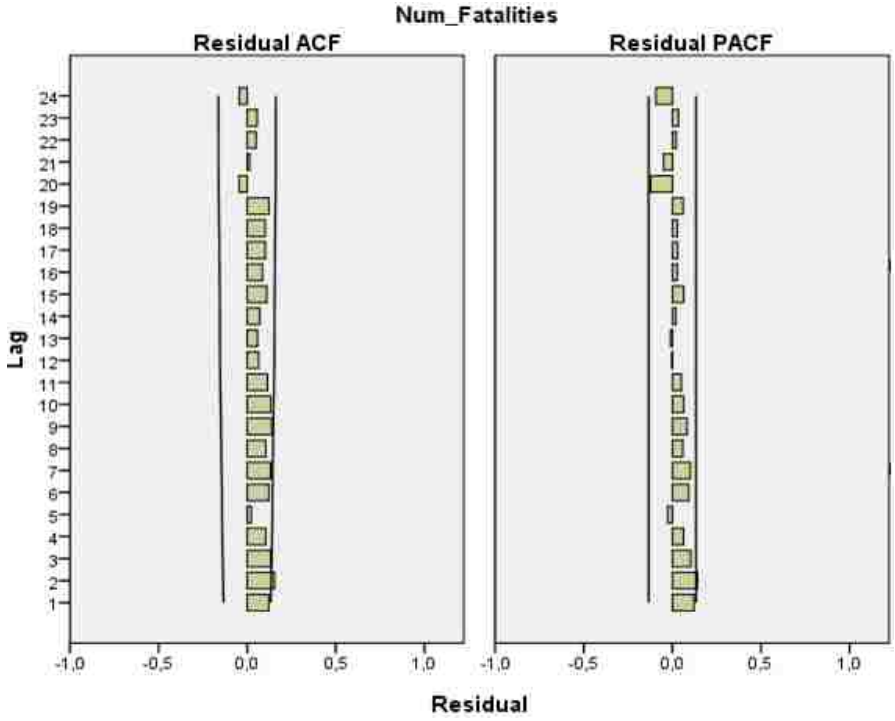


Figure 5.3 Residual ACF and PCAF of the Number of Fatalities.

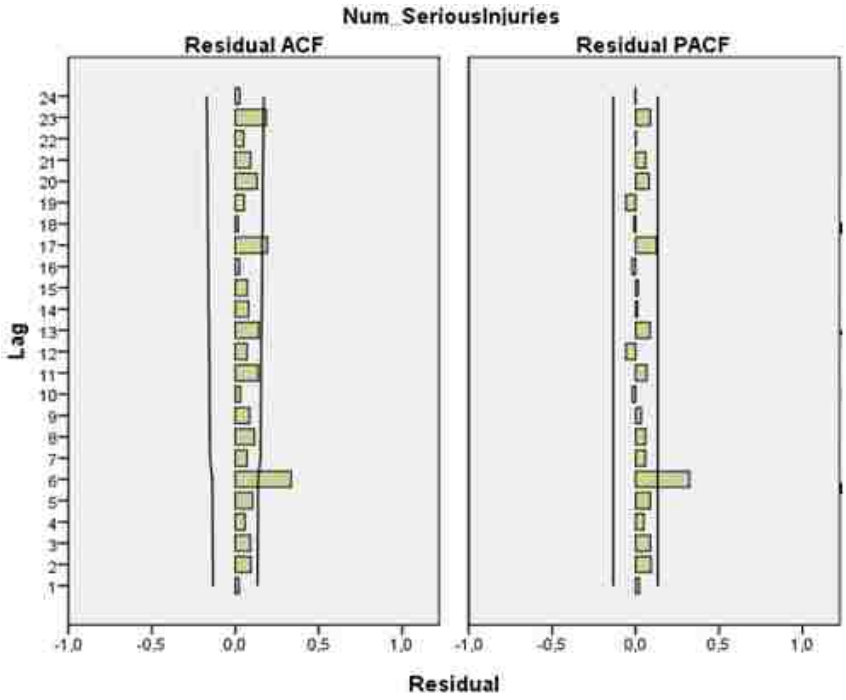


Figure 5.4 Residual ACF and PCAF of the Number of Serious Injuries.

The two time series datasets for the fatalities and serious injuries per 100 Million VMT, had 19 observed points of data (one point per year). Since this is data on an annual level, no seasonal factors were observed. Multiple combinations of ARIMA models with various *p*, and *q* values were applied and tested with non-seasonal differenced values.

A need was identified for differencing annual crash-rate data by checking stationarity. The plots for the ACF and PACF were used to identify *p, d,* and *q* values. Based on the goodness-of-fit measures, low MAPE and low RMSE values, the ARIMA (0, 1, 3) model was selected as having the best fit for forecasting fatalities and serious injuries, normalized with exposure variable VMT. Table 5.4 provides results of the goodness-of-fit for the various ARIMA models for fatalities and serious injuries per 100 Million VMT. Figure 5.5 and 5.6 illustrates the best fit ARIMA(0,1,3) model to forecast fatalities and serious injuries per VMT.

Table 5.4 Goodness-Of-Fit for the ARIMA Models

Stochastic Model/Performance Measures	Fatalities/ 100 Million VMT		Serious Injuries/100 Million VMT	
	MAPE	RMSE	MAPE	RMSE
ARIMA(0,1,0)	7.254	0.158	5.232	0.984
ARIMA(0,1,1)	7.105	0.162	5.054	1.009
ARIMA(0,1,2)	7.209	0.168	4.679	0.907
ARIMA(0,1,3)	6.587	0.167	4.763	0.928

The SARIMA(0,0,5)$(0,1,1)_{12}$ model was the optimal choice to forecast the number of fatalities and serious injuries for Nevada. The data used for the model showed an excellent number of observations. In addition, the validation datasets for the years 2013 and 2014 were checked against the forecasted data. In 2013, 268 data points were observed versus 269 forecasted fatalities. In 2014, 273 were observed versus 284 forecasted fatalities.

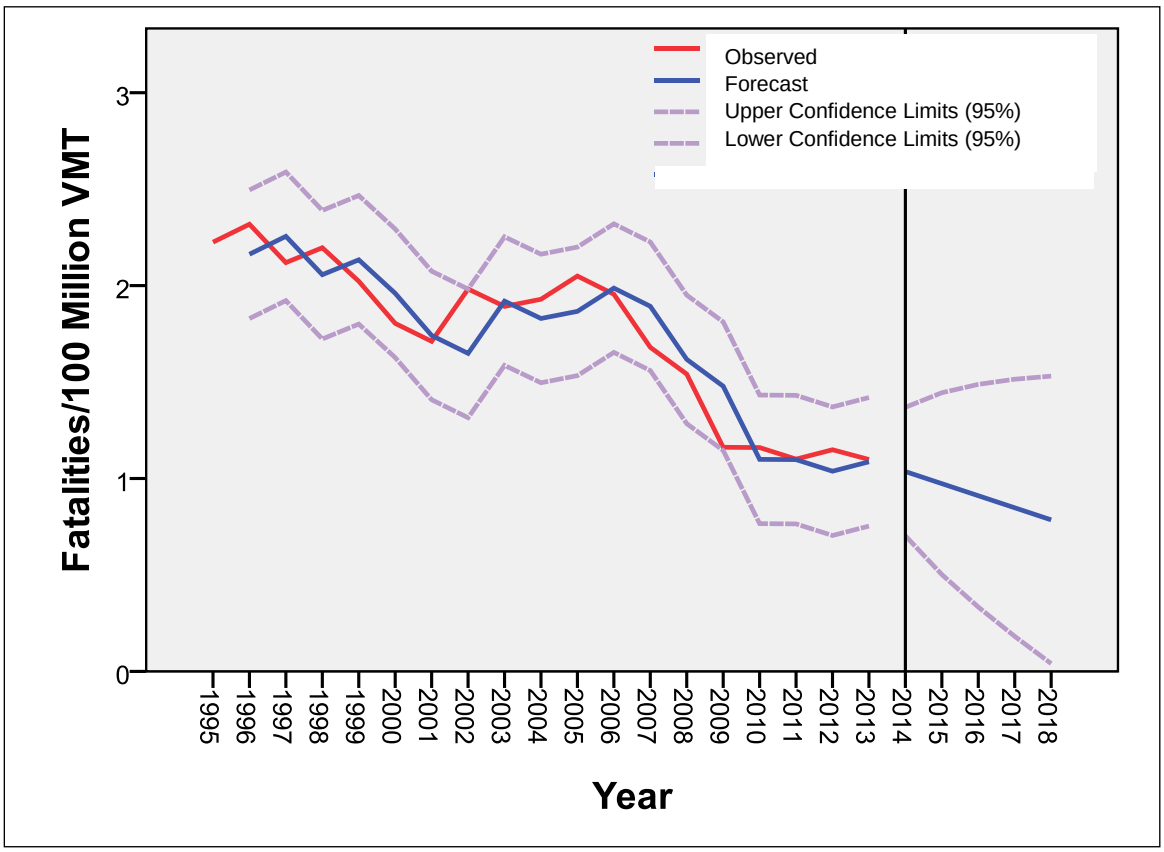


Figure 5.5 Forecast of the Rate of Fatalities using the ARIMA(0,1,3) model.

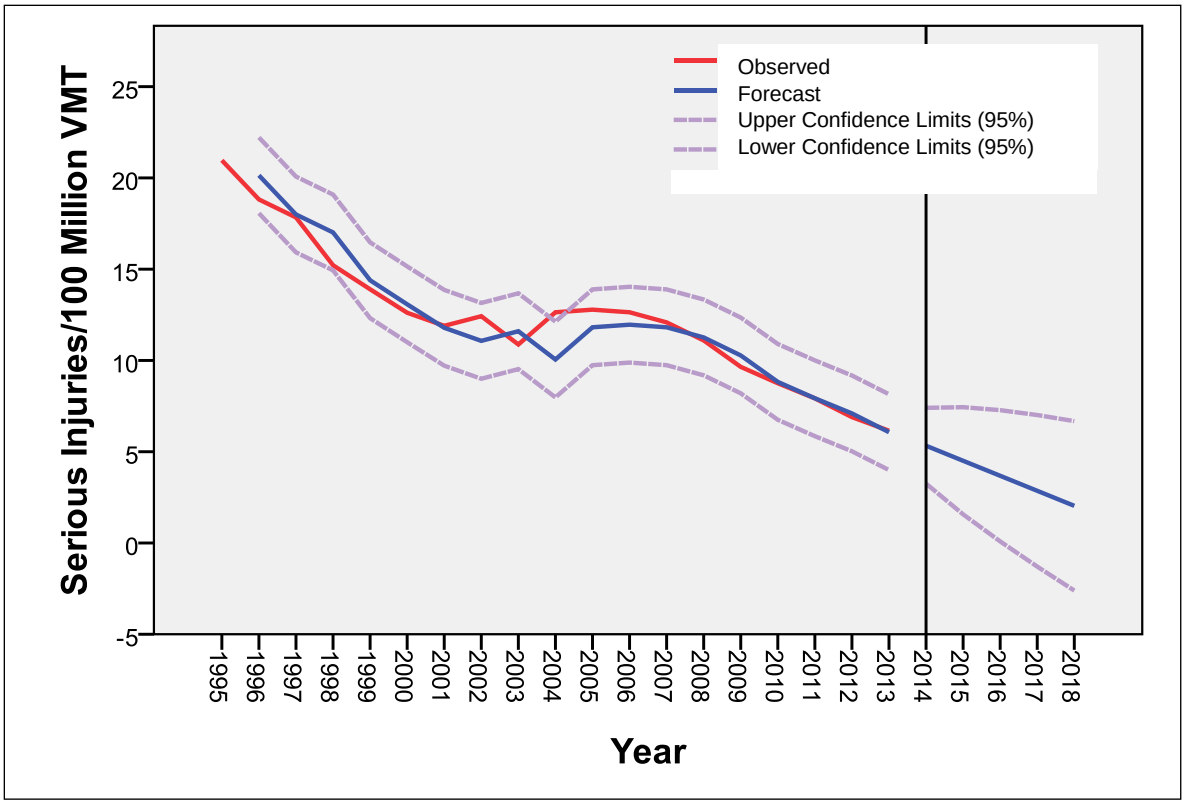


Figure 5.6 Forecast of the Rate of Serious Injuries using the ARIMA(0,1,2) model.

Similar results were obtained for the number of serious injuries. Based on the statistical results, the SARIMA model developed in this study was highly reliable in forecasting the number of fatalities and serious injuries in Nevada.

The developed models were included in the Business Intelligence framework using the methodology described in Chapter 3, Section 3.2. The crash data was accessed from the source crash database, NCATS. Time series models were coded using Oracle R and embedded in OBIEE. Dashboard was created to display the results as shown in Figure 5.7.

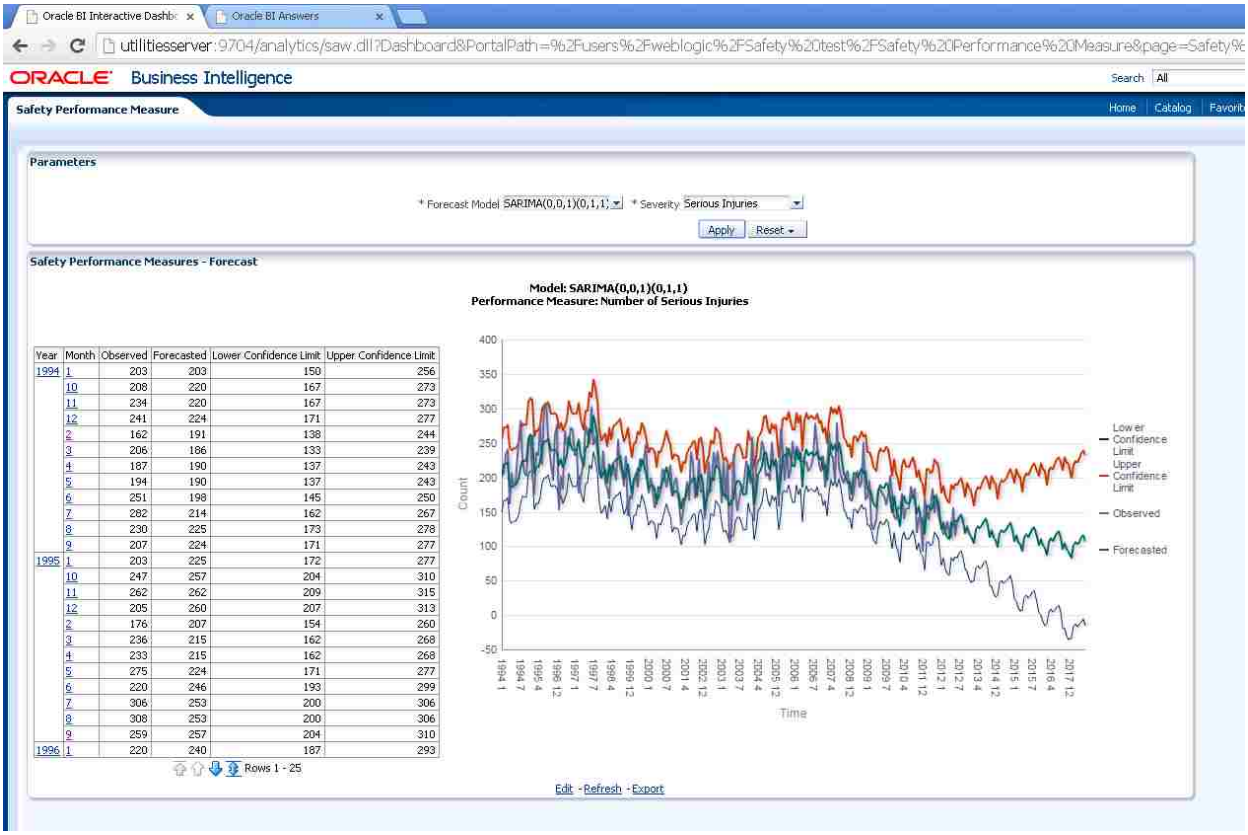


Figure 5.7 Dashboard Illustrating the Forecast of the Number of Serious Injuries using Time Series Models.

5.4 Conclusions

This research aimed to forecast traffic-safety performance measures using actual crash data. These forecasts could be used to determine targets for future safety-improvement programs and policies, when compared to existing methods currently in practice. From the perspective of a state DOT, predicting the number of fatalities and serious injuries is of significant importance to meet the requirements of MAP-21.

Historically, aspirational methods or models based on five years moving average projections have been used by public sector agencies. However, these methods and models may not be in line with actual data or adequate statistics. In many cases, such projections have led to grossly overestimating or underestimating traffic safety, leading to excess expenditures on safety

improvement programs or the loss of valuable lives. Decision-makers need to have access to robust crash-forecast models that enable them to prioritize and implement realistic and economically viable safety policies and programs. In this research, deterministic and stochastic models were developed to address this requirement. The best model specification was obtained using RMSE and MAPE as the goodness-of-fit.

In the case of deterministic models, the Winter-additive model for seasonal data and the Damped-trend model for non-seasonal data provided adequate forecasts. In the case of stochastic models, although the ARIMA model had an acceptable goodness-of-fit for non-seasonal data, the sample size (19 data points) is small according to Wei (Wei, 1990) who suggested a sample size with a minimum of 50 observations. The absence of large datasets likely will preclude an appropriate estimation when using the ARIMA (*p, d, q*) model (Wei, 1990; Harvey, 1990).

The SARIMA model was determined to be the best for use with Nevada data. Specifically, the stochastic SARIMA(0,0,5)(0,1,1) model seemed to be an improved model with a preferred fit for predicting the number of fatalities and serious injuries over a five-year horizon. This SARIMA model could be an appropriate statistical tool to predict fatalities and serious injuries, and an excellent asset for state DOT requirements of MAP-21. The methodologies used in this study were generalized and could be applied to time-series crash datasets in other states.

Further research that is recommended includes applying additional exposure and safety-intervention variables to the models in order to capture the effect of those variables on the forecasted number of fatalities and serious injuries.

CHAPTER 6

CONCLUSIONS, CONTRIBUTIONS AND RECOMMENDATIONS

6.1 Summary and Conclusions

This dissertation proposes a comprehensive business intelligence framework for network-level traffic safety analysis. The proposed framework addresses both methodological and practical barriers to enable practitioners to use theoretically sound methods. This research focused on comprehensive network screening analysis and regional level forecasts of performance measures to reduce fatalities and serious injuries for performance based traffic safety.

To facilitate quality control and benefit from existing resources, results were compared with those obtained using Safety Analyst, a state-of-the-art software. Algorithms from the *HSM* were reimplemented and used to perform network screening. A comprehensive database system was developed to provide data to multiple applications for traffic safety engineering and other potential needs. Furthermore, a methodology and guidelines are provided to develop similar databases from the existing, readily available data sources of state DOTs and/or MPOs. In particular, the proposed database system has the capability to provide data to Safety Analyst, a state-of-the-art highway safety management software. Although Safety Analyst provides tremendous analysis capabilities, few agencies take advantage of these capabilities because significant data needs, complex development of the required inputs is involved, and lack of experience and knowledge in creating the inputs as well as using the software.

The proposed database system, along with its data management and visualization tools, provides significant support to circumvent these barriers. To test the proposed system and tools,

data from Clark County was used to develop a database and perform the analyses with Safety Analyst. Specifically, this study examined the results from two case studies. The first case study, which identified sites having a potential for safety improvements with respect to fatal and all injury crashes, included all roadway elements and used default and calibrated Safety Performance Functions (SPFs). The second case study identified sites having a potential for safety improvements with respect to fatal and all injury crashes, specifically regarding intersections; it used default and calibrated SPFs as well. Guidelines were provided about the selection of a particular network screening type or performance measure for network screening. The proposed system enables the use of state-of-the-art traffic safety tools to support the development of federal requirements as well as to develop better traffic safety solutions for existing and emerging problems. The results obtained from this analysis are used as bench mark for further tasks.

Currently, practitioners are choosing easy-to-implement legacy methodologies which may lead to identification of incorrect sites with safety needs, thus resulting in inefficient traffic safety management. Traditionally, separate tools are used to integrate, process, and manage data; for modeling analysis; and to visualize the results. However, this traditional approach may result in data processing replication; it requires technical knowledge and consumes significant time. Hence, this research aimed to develop a comprehensive BI framework to enable practitioners to use existing as well as new, proposed here, theoretically sound methodologies with less effort, knowledge and time. The proposed framework was developed to address barriers associated with data integration, management, and visualization for the implementation of theoretically sound methodologies such as those in the *HSM* and expansions proposed here. The outcome is a single framework that accesses the data from a source, integrates and manages the data, processes analytical models, and provides results by means of a web-based interface. With the proposed

framework, network screening algorithms from the *HSM* were implemented. Network screening results produced by this framework were verified using results from Safety Analyst. Intuitive dashboards created for analysis and results enable to use the proposed framework with little effort.

A corridor level network screening approach based on the Empirical Bayes method was proposed and implemented for Fixed Corridor and Corridor Search algorithms. In contrast to the existing methods for corridor screening, expected crash frequencies were used, instead of observed crash frequencies or rates, to address the regression-to-the-mean bias for selecting corridors with potential for safety improvements. Top ranked corridors obtained using the proposed approach for corridor-level network screening were compared with ranked corridors using rate and frequency methods. The order of ranks of the corridors are completely different as a consequence of using a theoretically sound approach.

To improve the estimation of predicted crash frequency, SPFs for crash severity were estimated considering simultaneously crash patterns and associated explanatory characteristics. The objective was to minimize the estimation error by considering multiple SPFs (or clusters) rather than a single SPF for a given site and/or crash type. A mathematical program was formulated to assign similar crash sites into clusters and simultaneously seek sets of parameter values for the corresponding SPFs that maximize the probability of observing the data. A solution algorithm was developed using simulated annealing coupled with maximum likelihood estimation. Two site subtypes from Clark County, Nevada, were analyzed: 1) roadway segments for urban multi-lane divided arterials, and 2) urban 4-leg signalized intersections. The results obtained from the proposed approach were compared with the results obtained using a single SPF. Models were validated and results using multiple SPFs

improved the predicted number of crashes. In addition, the proposed approach to estimate SPFs improved the network screening results.

For performance based traffic safety analysis, to reduce fatalities and serious injuries, this research aimed to forecast traffic-safety performance measures using actual crash data. These forecasts could be used to determine targets for future safety-improvement programs and policies including MAP-21, when compared to existing methods currently in practice. In this research, deterministic and stochastic models were developed to address this requirement. The SARIMA model was determined to be the best for use with Nevada data. This SARIMA model could be an appropriate statistical tool to predict fatalities and serious injuries, and an excellent asset for state DOTs to use to meet the requirements from MAP-21.

6.2 Research Contributions

The first contribution of this research is a theoretically sound methodology for corridor level network screening. The proposed methodology was developed, implemented, and tested using actual crash data from the state of Nevada. Subject to a corridor length and step size, the proposed methodology estimates expected crash frequency for all roadway elements including segments, intersections, and ramps, within the corridor. Corridors are searched and ranked for the entire network. The use of expected crash frequency for corridor-level network screening is in contrast to existing methods which use observed crashes or associated rates which are subject to regression to the mean bias.

For network screening analysis, predicted crash frequency is a key component in estimating the expected crash frequency. The second contribution of this research is a

methodology for the estimation of multiple SPFs for a site or crash subtype based on crash patterns and associated explanatory characteristics. Multiple SPFs for a single site or crash subtype are able to better capture crash patterns and use explanatory information. Clusters of data are created by grouping records based on crash patterns and associated characteristics. The proposed methodology uses clusterwise regression to assign crash sites into clusters and estimate the coefficients of the corresponding SPFs simultaneously. To the best of author's knowledge, no previous study has attempted to utilize clusterwise regression to estimate the regression parameters for SPFs. The results of network screening analysis using the proposed methodology indicated better results than those obtained from traditionally calibrated SPFs.

The third contribution of this research involved the investigation of forecasting performance measures for traffic safety using deterministic and stochastic time series models. Results indicated that the stochastic time series models are best to use with Nevada data. Current practice is not based on a statistically sound methodology but rather relies on simplistic approaches such as moving averages. With an accurate statistically validated forecast models, projections of traffic safety performance measures are not expected to be over or underestimated. With access to robust crash-forecast models, decision-makers can prioritize and implement realistic and economically viable safety policies and programs.

Finally, a practical contribution of this research is the development of a single web-based BI framework to access and integrate source data, generate theoretically sound traffic safety analyses and provide visualization capabilities. The proposed framework reduces the effort and advanced technical knowledge required to perform traffic safety analysis using theoretically sound methodologies including the ones developed as part of this research as well as those recommended by the HSM. The proposed framework was tested using data from Nevada. Most

agencies use similar source data to perform traffic safety analysis. Considering that data management is automated, each year, development cost and time are minimized to perform analysis and visualize the desired results.

6.3 Future Research Recommendations

The field of data science is evolving exponentially with new techniques. As this research explored BI concepts and frameworks for traffic safety analysis, several future research directions arise and could be investigated. These include:

1) The data management procedures and algorithms were embedded within the proposed BI framework. However, no effort was spent on optimizing scripts to save computer resources and running time.

2) To complete the traffic safety management process, along with proposed network screening analysis, diagnosis, countermeasure selection, economic

3) analysis and priority ranking, and countermeasure evaluation algorithms should be developed and included within the proposed framework. This will result in a comprehensive traffic safety management process.

4) The mathematical program for the estimation of SPFs utilizing clusterwise regression should be expanded to include only significant explanatory variables. This will include cluster specific significant explanatory variables which may result in more accurate parameters for SPFs than the ones determined using the methodology proposed in this research. In addition, the mathematical program can be improved to determine the optimum number of clusters as one of the decision variables.

5) The mathematical program used for the estimation of SPFs can be tested with various alternate objective functions such as the Bayesian Information Criteria, Total Absolute Bias and Sum of Absolute Residuals. This could produce better goodness of fit and parameter estimates.

6) Similar to alternate objective functions, various model equations for SPFs can be tested. In this research, a multiplicative power model equation was used. Some examples of other model equations include polynomial, logistic, Weibull, exponential, Hoerl, Sigma, and combinations of Weibull and linear. There is less guidance available on which model form to choose on which type of data. For the same data, estimating with different model functions may differ in the parameter estimates.

7) Currently, Negative Multinomial distribution has been chosen based on characteristics of the available data. The proposed framework could be expanded to perform exploratory data analysis (EDA). Based on the results of EDA, appropriate distribution such as Poisson, Negative Binomial, Zero- inflated Poisson, Zero-inflated Negative Binomial, hurdle models can be chosen.

8) For performance based traffic safety program, methods to forecast performance measures were provided in this research. Future research can include the development of a methodology to set realistic targets to reduce fatalities and serious injuries. Realistic targets should be set based on data-driven analysis. Historically, DOTs spent their monitory resources on various counter measures to reduce fatalities and serious injuries. Budget-to-cost (BC) ratio should be estimated for implemented countermeasures on various critical emphasis areas and countermeasure evaluation. With future estimated budgets available, selection of countermeasures should be optimized to provide large BC

ratios as well as high reduction of fatalities and serious injuries. With the selected countermeasures, estimation of reduction in fatalities and serious injuries should be accounted during the target setting process.

REFERENCES

Alluri, P. (2008). Assessment of Potential Site Selection Methods for use in Prioritizing Safety Improvements on Georgia Roadways, Master's Thesis report.

Alluri, P. (2010). *Development of Guidance for States Transitioning to New Safety Analysis Tools*. Doctoral Degree Dissertation.

Alluri, P. & J. Ogle. (2011). Road Safety Analysis: States' Current Practices and Their Future Direction. *91st Annual Meeting Transportation Research Board*, CD-ROM. Transportation Research Board of the National Academies, Washington, D. C.

Alluri, P., & Ogle, O. (2012). Road safety analysis in the United States. In *Transportation Research Record: Journal of the Transportation Research Board,* no. *2318,* 7-15.

American Association of State Highway and Transportation Officials. (2010). *Highway Safety Manual*. American Association of State Highway and Transportation Officials.

American Association of State Highway and Transportation Officials. (2011). Safety Analyst. Retrieved from http://www.safetyanalyst.org. Accessed on Nov 15, 2011.

AASHTOWare Safety Analyst v4.3.3. (2014). *Safety Analyst Newsletter.* Retrieved from http://www.aashtoware.org/Safety/Pages/Safety-Analyst-Newsletter.aspx

A Primer for Dynamic Traffic Assignment. (2010). Transportation Network Modeling Committee. Transportation Research Board. Retrieved from http://nextrans.org/ADB30/sites/default/files/dta_primer.pdf. Accessed Mar 15, 2011.

ArcGIS Web API.(2015). ArcGIS API for Javascript. Retrieved from https://developers.arcgis.com/javascript/latest/api-reference/index.html

Aylo, R. (2010). *Wave propagation in negative index materials*. (Electronic Thesis or Dissertation). University of Dayton.

Berenson, M. L., Krehbiel, T.C., & Levine, D.M. (2005). Confidence interval estimation. (10th Ed.), *Basic business statistics: Concepts and applications*. New Jersey: Prentice Hall.

Body, M., Miquel, M., Bédard, Y., & Tchounikine, A. (2002). A multidimensional and multiversion structure for OLAP applications. *DOLAP'02*, McLean, Virginia, USA, 2002.

Bolker, B. (2016). Tools for general maximum likelihood function – Package 'bbmle'. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/bbmle/bbmle.pdf

Box, G., Jenkins, G., Reinsel, G. (2008). Time series analysis: Forecasting and Control, 4th ed., New York: John Wiley& Sons.

Brusco, M. J., Cradit, J.D., Steinley, D., & Fox, G.L. (2008). Cautionary Remarks on the Use of Clusterwise Regression. *Multivariate Behavioral Research, 43*(1), 29-49.

Cadkin, J. (2002). Understanding dynamic segmentation-working with events in ArcGIS 8.2, ESRI. ArcUser October-December. Retrieved from http://www.esri.com/news/arcuser/1002/files/dynseg_2.pdf.

Carlsson, C., & Turban, E. (2002). DSS: Directions for the next decade. *Decision Support Systems, 33*(2), 105-110.

Center for Advanced Public Safety (CAPS). (2009). *CARE – Critical Analysis Reporting Environment*. Retrieved from http://caps.ua.edu/online_analysis.aspx.

Collins, N. E., Eglese, R.W., & Golden, B.L. (1988). Simulated annealing – An annotated bibliography. American Journal of Mathematical and Management *Science, 8*(3–4), 209-307.

Council, F. M., Harkey, D.L., Carter, D.L., & White, B. (2007). *Model Minimum Inventory of* 7 *Roadway Elements—MMIRE*. FHWA-HRT-07-046, Federal Highway Administration, McLean, VA.

Chen, H., Chiang, R.H.L., & Storey, V.C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165-1188.

Data Integration Primer. (Aug. 2001). Federal Highway Administration. U.S. Department of Transportation.

Devogele, T., Parent, C., Spaccapietra, S. (1998). On spatial database integration. *International Journal of Geographic Information Sciences*, *12*(4), 335-352.

Depaire, B., Wets, G., & Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention*, *40*, 1257-1266.

DeSarbo, W. S., Oliver, R.L., & Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear regression. *Psychometrika*, *54*(4), 707–736.

Dueker, K., & Butler, J.A. (1998). GIS-T enterprise data model with suggested implementation choices. *Journal of the Urban and Regional Information Systems Association*, *10*(1), 12-36.

Dupupet, C., Gray, D., Boyd-Bowman, P.C., & Testut, J. (2013). *Oracle data integrator 11g cookbook*. Packt Publishing, UK.

DynusT User's Manual. (2008). Retrieved from http://dynust.net/ wikibin/doku.php. Accessed Mar 02, 2009.

Federal Highway Administration. (2013). Map-21 Putting Performance into Action. *Safety Provisions in Moving Ahead for Progress in the 21st Century (MAP-21)*. Retrieved from http://www.fhwa.dot.gov/map21/safety_overview.cfm. Accessed July 31, 2013.

Federal Highway Administration (2015). *Highway safety improvement program reports*, U.S. Department of Transportation.

Freeman, T., & Tukey, J. (1950). Transformations Related to the Angular and the Square Root. *The Annals of Mathematical Statistics, 21*(4), 607-611.

Fridstrøm, L., Ifver,J., Ingebridtsen, S., Kulmala, R., & Thomsen, L. (1995). Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts. *Accident Analysis and Prevention, 27*(1), 1-20.

Gan, A., Haleem, K., Alluri, P., Lu, J., Wang, T., Ma, M., Diaz, C . (2012). Preparing Florida for Deployment of SafetyAnalyst for all roads. *Miami: Lehman Center for Transportation Research*, Florida International University.

Geedipally, S., & Lord, D. (2010). Investigating the effect of modelling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models. *Accident Analysis and Prevention, 42*(4), 1273-1282.

Geoprocessing with ModelBuilder, *ESRI ArcGIS*. Retrieved from http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#. Accessed April 2013.

Guo, J. Q., & Zheng. L. (2005). A modified simulated annealing algorithm for estimating solute transport parameters in streams from tracer experiment data. *Environmental Modelling & Software, 20*, 811–815.

Hall, J.P., Robinson, R., & Paulis, M.A. (2005). Enterprisewide Spatial Data Integration of Legacy Systems for Asset Management–The Case of the Illinois Department of Transportation. *Transportation Research Record: Journal of the Transportation Research Board, No.1917*, Transportation Research Board of the National Academics, Washington, D.C., 11–17.

Hamidi, A., Fontaine, M.D., & Demetsky, M.J. (2010). A Planning-Level Methodology for Identifying High-Crash Sections of Virginia's Primary System. *Virginia Transportation Research Council, Final Report* (No. VTRC 11-R4, 2010). Retrieved from http://www.virginiadot.org/vtrc/main/online_reports/pdf/11-r4.pdf.

Hauer, E. (1997). *Observational before-After Studies in Road Safety*. Pergamon Press Inc.

Hauer, E. (2004). Statistical Road Safety Modeling, *Transportation Research Record: Journal of the Transportation Research Board, No. 1974*, 799-808.

Hauer, E. (2015). The Art of Regression Modeling in Road Safety, New York, Springer.

Hauer, E., Harwood, D.W., Council, F.M., & Griffith, M.S. (2002). Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transportation Research Record: Journal of the Transportation Research Board, No.1784*, Transportation Research Board of the National Academics, Washington, D.C., 126–131.

Hamidi, A., Fontaine, M.D., & Demetsky, M.J. (2010). A planning-level methodology for identifying high-crash sections of Virginia's primary system. *Virginia Transportation Research Council – Research Report* (No. FHWA/VTRC11-R4).

Highway Safety Improvement Plan. HSIP Noteworthy Practice Series – HSIP Project Identification. Retrieved from http://safety.fhwa.dot.gov/hsip/resources/fhwasa1102/id.pdf.

Highway Safety Manual. (2010). American Association of State Highway and Transportation Officials.

Highway Performance Monitoring System. (2010). Federal Highway Administration. Retrieved from http://www.fhwa.dot.gov/policyinformation/hpms/fieldmanual/chapter1.cfm. Accessed March 10, 2012

ITT Corporation. (2011). *Safety Analyst user's manual*. Colorado Springs, CO.

Inmom, W.H. (3rd Ed.). (2002). *Building the data warehouse*. New York, NY: John Wiley & Sons, Inc.

iTRANS Consulting Ltd, and Human Factors North INC. (2003). *New Approaches to Highway Safety Analysis*.

Karlaftis, M.G., & Tarko, A.P. (1998). Heterogenity considerations in accident modeling. *Accident Analysis and Prevention, 30*(4), 425-433.

Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Khan, G., Santiago-Chaparro, K.R., Chiturri, M., & Noyce, D.A. (2010). Development of Data Collection and Integration Framework for Road Inventory Data. *Transportation Research Record: Journal of the Transportation Research Board, No. 2160*, Transportation Research Board of the National Academics, Washington, D.C., 29–39.

Khanal, I. (2014). Development of a Visualization System for Highway Safety Management using Safety Analyst, Master's Thesis.

Kimball, R., & Merz, R. (2000). *The data webhouse toolkit*. New York, NY: John Wiley &Sons, Inc.

Kweon, Y., & Lim, I. (2012). Appropriate regression model types for intersections in SafetyAnalyst. *Journal of Transportation Engineering*, American Society of Civil Engineers, *138*(10), 1250-1258.

Lau, K., Leung, P., & Tse, K. (1999). A mathematical programming approach to clusterwise regression model and its extensions. *European Journal of Operational Research*, no. 116, 640-652.

Li, H., Jou, J.P., Zhao, X.F., Zhou,W., Li, H., Zhou, Z., & Yang, Y. (2006). Structural health monitoring system for the Shandong Binzhou Yellow River highway bridge. *Computer-Aided Civil and Infrastructure Engineering*, *21*(4), 306-317.

R-Language. (2011). *A Mailing list for language researchers who use R statistical programming language for modeling and data analysis*. Retrieved from https://mailman.ucsd.edu/pipermail/ling-r-lang-l/2011-August/000282.html.

Lord, D., Guikema, S., & Geedipally, S.R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention*, *40*(3), 1123–1134.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291-305.

Lu, J., Haleem, K., Alluri, P., Gan, A., & Liu, K. (2013). Developing local safety performance functions versus calculating calibration factors for SafetyAnalyst applications: A Florida case study. *Safety Science*, *65*, 93-105.

Lu, H., Huang, S., Li, Y., & Yang, Y. (2014). Panel Data Analysis via variable selection and subject clustering. *Data Mining for services, studies in Big Data 3,* Springer-Verlag.

Luo, Z., & Chou, E.Y.J. (2006). Pavement condition prediction using clusterwise regression. *Transportation Research Record: Journal of the Transportation Research Board, No. 1974*, 70–77.

McDermid, D., and Taft, M. (2014). *Oracle R Enterprise User's Guide*, Release 1.4. Oracle.

Miaou, S.P., Song, J.J., & Mallick, B.K. (2003). Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics*, *6*(1), 33–57.

Milton, J., Shankar, V., & Mannering, F. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis and Prevention, 40*(1), 260-266.

Ming, Q., & Lei, S. (2010). Comprehensive transportation Information Platform based CORBA Framework and XML data Exchange Technology. *ICCTP 2010: Integrated Transportation Systems – Green, Intelligent and Reliable Conference Proceedings*, 1218-1225.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., & Ukkusuri, S.V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. *Safety Science, 54*, 27-37.

Montella, A. (2010). A Comparative Analysis of Hotspot Identification Methods. *Accident Analysis & Prevention, 42*(2), 571-581.

National Cooperative Highway Research Program (NCHRP) Report 666. (2010). Target-Setting methods and Data management to support performance based Resource Allocation by Transportation Agencies, Transportation Research Board of the National Academics.

National Highway Traffic Safety Administration. (2013). *Budget Estimates Fiscal Year 2013.* Retrieved from http://www.nhtsa.gov/Laws+&+Regulations/NHTSA+Budget+Information. Accessed May 30, 2013.

National Highway Traffic Safety Administration. (2013). *Early Estimates of Motor vehicle Traffic Fatalities in 2012* . Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/811741.pdf. Accessed May 30, 2013.

National Highway Traffic Safety Administration. *Model Minimum Uniform Crash Criteria*. Retrieved from http://www.mmucc.us/2008MMUCCGuideline.pdf. Accessed Dec 11, 2011.

Nemmers, C.J., Vap, D., & McDonald, T.J. (2008). Effectiveness of safety corridor programs, report on tasks 1–3. *Midwest Transportation Consortium*, (No. MTC Project 2007-08).

Nevada Department of Transportation. (2016). *Traffic Safety Engineering*. Retrieved from http://www.nevadadot.com/About_NDOT/NDOT_Divisions/Planning/Traffic_Safety_Engineering/Traffic_Safety_Engineering.aspx

Ogle, J. H. (2007). Technologies for Improving Safety Data: A Synthesis of Highway Practice. In *Transportation Research Record: Journal of the Transportation Research Board, No. 367*, Transportation Research Board of the National Academies, Washington DC.

O'Packi, P., Dubois, R., Armentrout, N., & Bower, S. (2000). Maine's Approach to Data Warehousing for State Departments of Transportation. *Transportation Research Record: Journal of the Transportation Research Board, No. 1719*, Transportation Research Board of the National Academics, Washington, D.C., 227–232.

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.

Park, M. W., & Kim, Y.D. (1998). A systematic procedure for setting parameters in simulated annealing algorithms. *Computer Operational Research, 25*(3), 207–217.

Paz, A., Khanal, I., Veeramisti, N., Baker, J., & Belmonte, L. (2014). Development of a visualization system for Safety Analyst. *Transportation Research Record: Journal of the Transportation Research Board, no. 2460*, 176-185.

Paz, A., Molano, V., Martinez, E., Gaviria, C., & Arteaga, C. (2015a). Calibration of Traffic Flow Models Using a Memetic Algorithm. *Transportation Research Part-C: Emerging Technologies, 55*, 432-443.

Paz, A., Molano, V., & Sanchez, M. (2015b). Holistic Calibration of Microscopic Traffic Flow Models: Methodology and Real World Application Studies. *Engineering and Applied Sciences Optimization: Dedicated to the memory of Professor M.G. Karlaftis. 38*(1). Springer International Publishing. ISBN 978-3-319-18320-6.

Paz, A., Veeramisti, N., Khanal, I., Baker, J., & de la Fuente-Mella, H. (2015c). Development of a comprehensive database system for Safety Analyst. *The Scientific World Journal, 2015*, 2015.

Pendyala, R. M., & Oak Ridge National Laboratory. (2008). *Development of GIS-Based Conflation Tools for Data Integration and Matching: Final Report.* Florida Department of Transportation, Tallahassee, Fla., 2002. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C.

Pine, M., Sonneborn, M., Schindler, J., Stanek, M., Maeda, J., & Hanlon, C. (2012). Harnessing the power of enhanced data for healthcare quality improvement: Lessons from a Minnesota hospital association pilot project. *Journal of Healthcare Management, 57*(6), 406–418.

Poggi, J., & Portier, B. (2011). PM10 forecasting using clusterwise regression. *Atmospheric Environment, 45*, 7005-7014.

Pulugurtha, S.S., Krishnakumar, V.K., & Nambisan, S.S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Journal of Accident Analysis and Prevention, 39*(4), 800-811.

Qin, X., Sultana, M., Chitturi, M., & Noyce, D. (2013). Developing a truck corridor crash severity index. *Transportation Research Record: Journal of the Transportation Research Board*, no. *2386*, 103-111.

Riccardi, G. (2002). *Database Management with Web Site Development Applications*. Prentice Hall.

Rittman, M. (2013). *Oracle Business Intelligence 11g Developer's Guide*. USA: The McGraw-Hill Companies.

Román-Román, P., Romero, D., Rubio, M.A., & Torres-Ruiz, F. (2012). Estimating the parameters of a Gompertz-type diffusion process by means of Simulated Annealing. *Applied Mathematics and Computation*, *218*(9), 5121-5131.

Roshan S. B., Jooibari, M.B., Teimouri, R., Asgharzade-Ahmadi, G., M. Falahati-Naghibi, M., & Sohrabpoor, H. (2013). Optimization of friction stir welding process of AA7075 aluminum alloy to achieve desirable mechanical properties using ANFIS models and simulated annealing algorithm. *International Journal of Advanced Manufacturing Technology*, *69*(5-8), 1803–1813.

Rose, J., Klebsch, W., & Wolf, J. (1990). Temperature measurement and equilibrium dynamics of simulated annealing placement. *IEEE Transactions on Computer Aided Design*, *9*(3), 253–259.

Safe Accountable Flexible Efficient Transportation Equity Act: A Legacy for Users. A Summary of Highway provisions in SAFETEA-LU, 2005. Retrieved from https://www.fhwa.dot.gov/safetealu/summary.htm.

Safety Analyst User Manual. (2011). ITT Corporation, Colorado Springs, CO.

Sasidharan, L., Wu, K., & Menendez, M. (2015). Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis and Prevention*, *85*, 219-228.

Selim, S. Z., & Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition, 24*(10), 1003–1008.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461-464.

Srinivasan, R., Carter, D., & Bauer, K. (2013). Safety performance function decision guide: SPF calibration vs SPF development. *Federal Highway Administration – Office of Safety Report*, Washington, DC.

Shimko, G. & Walbaum, C. (2010). City of Edmonton traffic safety strategy. *Road Safety Session of the 2010 Annual Conference of the Transportation Association of Canada*, Halifax, Nova Scotia.

Shugan, S. (2006). Editorial: errors in the variables, unobserved heterogeneity, and other ways of hiding statistical error. *Marketing Science*, 25(3), 203–216.

Song, E., Yin, S., & Ray, I. (2007). Using UML to model relational database operations. *Computer Standards and Interfaces*, *29*(3), 43-354.

Spath, H. (1979). Clusterwise linear regression. *Computing*, *22*(4), 367–373.

Tarko, A.P., Li, M., Romero, M., & Thomaz, J. (2014). A systematic approach to identifying traffic safety needs and intervention programs for Indiana: Volume I – Research report. In *Joint Transportation Res3earch Program Publication*, no. FHWA/IN/JTRP-2014/03.

Tjandra, S.A. (2014). Business Intelligence system for traffic data integration: linking roadway, collision and traffic flow data to improve traffic safety. *NATMEC 2014 Improving Traffic Data Collection, Analysis, and Use*. Chicago, Illinois.

Tremblay, M., Hevner, A.R., & Berndt, D.J. (2012). Design of an information volatility measure for health care decision making. *Decision Support Systems, 52*(2), 331-341.

Turban, E., & Walls, J.G. (1995). Executive information systems – a special issue. *Decision Support Systems,14*, 85-88.

Ulfarsson, G.F., & Shankar, V.N. (2003). An accident count model based on multi-year cross-sectional roadway data with serial correlation. *Transportation Research Record: Journal of the Transportation Research Board, No. 1840*, 193-197.

Volonino, L., Watson, H.J., & Robinson, R. (1995). Using EIS to respond to dynamic business conditions. *Decision Support Systems, 14*(2),105-116.

Vonderohe, A., Chou, C., Sun, F., & Adams, T. (1998). A generic data model for linear referencing systems. *Research Results Digest Number 218*, National Cooperative Highway Research Program, Transportation Research Board of the National Academics, Washington, D.C.

Wang, H., Li, A.Q., Tong, G., Tao, T. (2014). Establishment and application of the wind and structural health monitoring system for the Runyang Yangtze River Bridge. *Shock and Vibration, 2014,* Article ID 421038.

Wang, C., Quddus, M.A., & Ison, S.G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention, 43*, 1979-1990.

Washington, S.P., Karlaftis, M.G., & Mannering, F.L. (2nd Ed.) (2010). Statistical and econometric methods for transportation data analysis. Boca Raton, FL: Chapman Hall/CRC.

Wellner, A., & Qin, X. (2011). GIS-based highway safety metrics implementation and evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, no. *2241*, 1-9.

Wikipedia. *Database schema.* Retrieved from http://en.wikipedia.org/wiki/Database_schema. Accessed May 02, 2012.

Wu, Y.-J., Wang, Y., & Qian, D. (2007). A Google-Map-Based arterial traffic information system. *Proc., ITSC 2007: IEEE Intelligent Transportation Systems Conference,* Seattle, Washington, 968–973.

Xiao Qin, A. W. (2011). Development of safety screening tool for high risk rural roads in South Dakota. *Civil Laboratory for Operations and Safety Engineering in Transportation*, South Dakota State University.

Xie, G., & Hoeft, B. (2014). 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*, 23-34.

Zhao, Y., Tian, Z.Z., & Reider, C. Screening urban road networks for corridors with promise. *94th Transportation Research Board Annual Meeting*, Washington D.C.

Ziliaskopoulos, A.K. & Waller, S.T. (2000). An Internet based geographic information system that integrates data, models and users for transportation applications. *Transportation Research Part C: Emerging Technologies*, *8*(1-6), 427–444.

.

CURRICULAM VITAE

Graduate College
University of Nevada, Las Vegas

Naveen Kumar Veeramisti

Degrees:
Bachelor of Engineering, Civil Engineering, 2001
University of Madras, India

Master of Science, Civil and Environmental Engineering, 2007
University of Nevada, Las Vegas

Publications:
Alexander Paz, N. Veeramisti, R. Khaddar, and H. de la Fuente-Mella, (2015). Traffic and Driving Simulator Based on Architecture of Interactive Motion, The Scientific World Journal, Vol. 2015, Article ID 340576, 9 pages, 2015. doi: 10.1155/2015/340576.

Alexander Paz, N. Veeramisti, I. Khanal, J. Baker, and H. de la Fuente-Mella (2015). Development of a Comprehensive Database System for Safety Analyst, The Scientific World Journal, Vol. 2015, Article ID 636841, 14 pages, 2015. doi:10.1155/2015/636841.

Alexander Paz, I. Khanal, N. Veeramisti, J. Baker, and L. Belmonte (2014). Development of a Visualization System for Safety Analyst. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2460, pp176-185.

Alexander Paz, A. Nordland, N. Veeramisti, and A. Khan (2014). Assessment of Economic Impacts of a VMT Fee for Passenger Vehicles in Nevada. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2450, pp26-35.

Dissertation Title: A Business Intelligence Framework for Network-level Traffic Safety Analyses.

Dissertation Examination Committee:
Advisor, Alexander Paz, Ph.D., P.E.
Graduate Faculty Representative, Brendan Morris, Ph.D.
Committee Member, Mohamed Kaseko, Ph.D.
Committee Member, Moses Karakouzian, Ph.D., P.E.
Committee Member, Hualiang Teng, Ph.D.