TRANSPORTATION SAFETY ANALYTICS WITH STATISTICAL MACHINE LEARNING

by

Zuoyu Miao

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF SYSTEMS AND INDUSTRIAL ENGINEERING

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by ***Zuoyu Miao***, titled ***Transportation Safety Analytics with Statistical Machine Learning*** and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

K. Larry Head Date: (***11/20/2018***)

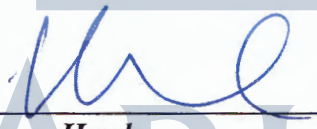
Yao-Jan Wu Date: (***11/20/2018***)

Jian Liu Date: (***11/20/2018***)

Lingling An Date: (***11/20/2018***)

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.



K. Larry Head Date: (***11/20/2018***)
Professor
Systems and Industrial Engineering

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor K. Larry Head. He is the most important person in my Doctoral life. I could not accomplish the Doctoral study without his support. He generously gave me the precious second chance to continue my study at University of Arizona. He encouraged me to explore my interests and helped me develop ideas about my research. He read and edited my manuscript drafts and polished them, making them published. He is the greatest professor I have ever worked with, with great personality and academic achievements. It is my great honor to complete my study and research under his guidance.

Much gratitude goes to my committee members: Dr. Yao-Jan Wu, Dr. Jian Liu, and Dr. Lingling An who have been supportive and encouraging with constructive recommendations for improving this dissertation.

Table of Contents

List of Figures

Chapter 3 Figures

Chapter 4 Figures

Chapter 5 Figures

Chapter 6 Figures

Chapter 7 Figures

List of Tables

Chapter 3 Tables

Chapter 4 Tables

Chapter 5 Tables

Chapter 6 Tables

Abstract

Transportation safety has been a topic of high priority for a long time, especially for the society's public service systems, e.g. the freight transportation system and public commuting system. With the development of modern transportation technologies, complex products have been designed and are being used by the public, such as the connected and autonomous vehicles. The potential risks brought by safety uncertainty of the transportation system is of public's concerns. Traffic accidents are among the top concerns. For example, in the past decade, there are approximately 100,000 traffic accidents in Arizona and 6,000,000 accidents nationwide every year. Some of them may incur very huge loss and some of them were observed to happen very often. Fortunately, with recent progress in Internet of Things (IoT), big data and machine learning technologies, there are new methods to analyze problems existing in the transportation safety analytics.

This dissertation investigates several important transportation safety issues. Utilizing multiple sources of data, new methods for analyzing accident severity and frequency are developed with consideration of the unique characteristics of accident data. In the accident severity analysis, a new method is proposed and applied in the imbalanced multi-class classification problem. Significant safety factors are identified and counter measures are proposed for improving safety. In the accident frequency analysis, a new quantitative method of predicting and assessing accident risks is proposed that takes into account of the relationship between travel time reliability and crash frequency.

Besides analyzing historic traffic accidents, this dissertation explores the utilization of connected vehicle data for improving safety. Machine learning techniques are used to predict a vehicle's trajectory so that potential conflicts and crashes can be identified. This is a powerful new approach

to analyzing vehicle data, but can also pose a risk to privacy and security. To investigate the risk, several supervised learning techniques are used to re-identify vehicles from vehicle data used for prediction. In a connected vehicle environment, connected vehicles have anonymous identifications. Using partial trajectory information from different intersections, the same connected vehicles are shown to be re-identifiable with high accuracy using machine learning techniques. Different factors that impact on the re-identification accuracy are evaluated within a designed experiment. For the purpose of predicting connected vehicle's future trajectory, deep learning techniques are applied with neural network structures with considerations of vehicle information, intersection signal and phase information and surrounding dynamic information. The proposed method is validated and compared with the classic object tracking algorithm and the naïve deep learning method.

1 Introduction

1.1 Background

Artificial Intelligence (AI) technologies are leading revolutions, including the one in the transportation industry. Changes are constantly made in infrastructure facilities and transportation tools, redefining people's relations with vehicles and generating huge amounts of transportation data. New perspectives of resolving hard transportation problems are becoming possibilities and proven to be more effective. Though machine learning technologies can be dated back long ago, the progresses made in the last decade are significant for understanding, interpreting and widely implementing such concepts and techniques. The recently emerged deep learning technologies provides even more options for solving very complex problems which were computationally unaffordable 10 years ago. This is primarily due to the idea of Back Propagation training algorithms (LeCun et al., 1988) and the fast development of computation hardware (e.g. Graphics Processing Unit). Meanwhile, due to the vast deployment of Internet of Things (IoT) devices (e.g. Connected and Autonomous Vehicle systems) and database facilities, it is easier to collect and store very large datasets. The combination of AI and Big Data technologies will definitely benefit both transportation industry and academia. This dissertation will focus on how those technologies can contribute to transportation safety issues in accident analysis and a connected vehicle environment.

Highway crash accident analysis is generally considered to include 2 aspects, the accident severity and frequency. Severity analysis is aiming at assessing the outcome of an accident, gaining understanding of contributing factors and deriving strategies for prevention and mitigating potential accident severity. The Highway Safety Information System (HSIS) Zhu and Srinivasan

(2011b) database quantifies the extent of injury into a numerical scale ranging from 1 to 5 (No Injury, Possible Injury, Non-incapacitating Injury, Incapacitating Injury and Fatal Injury). The other similar classification coding scheme uses KABCO (USDOT, 2003) to represent injury severities, though definitions for KABCO are different across states. Frequency analysis is to address the problem of accident risks. Usually the risk analysis is associate with the highway segment length and annual average daily traffic (AADT). The outcome is treated as an important safety measure performance for network screening, countermeasure comparison and project evaluation (Brimley et al., 2012).

In a connected vehicle environment, vehicles can communicate with nearby vehicles (V2V) through Onboard Units (OBU) devices and with nearby infrastructures (V2I) through Roadside Units (RSU). This wireless communication technology is called Dedicated Short Range Communication (DSRC) which operates at 5.9GHz spectrum (Antonucci et al., 2004). In this framework, vehicle data (GPS location, speed, acceleration, phase, etc.) can be collected every 100ms. The intersection each phase's status information is also available to users. The variety and number of data is potentially creating more opportunities for making the traffic running faster and safer.

1.2 Research Topic Statement

Statistical machine learning techniques are providing new perspectives and insights for transportation safety analysis. Two major aspects of this research are focused on highway truck crash accidents and connected vehicle trajectories. The specific research interests are presented below:

- *Highway Truck Crash Accident Severity Analysis*

Trucks, as the major transportation tool, has helped improved people's lives by moving goods all over the country and the number of them is increasing over time. Truck drivers are considered likely to bear higher risks of suffering serious injuries and fatalities than other type vehicle drivers and passengers (USDOT, 2005). The characteristics of truck driver and passenger's injury severity should be intuitively visualized and statistically validated, emphasizing the difference from other type vehicle related injury severity. An ideal analysis model is capable of considering a large number of potential factors (vehicle, roadway, weather, human, etc.) and identify critical ones. Factor influences on injury severity outcomes should also be quantified. Based on selected factors, corresponding strategies and measures of mitigating injury severity will be investigated and discussed.

- *Highway Truck Crash Accident Frequency Analysis*

Highway freight corridor's safety performance is evaluated by the probability of having crash accidents and the mobility performance is evaluated by the travel time reliability measures (e.g. buffer time). It is generally known that crash accidents would lead to potential congestion and increase the uncertainty of travel time. But the impacts of travel time uncertainty on crash risk are not well understood in past research literature and conclusions vary among different scenario settings. It is desired that a risk analysis model could consider whole corridor's travel time status as well as other factors (e.g. weather information). It is possible to identify critical corridor segments where congestion has significant impacts on accident risks and traffic management countermeasures can be exerted. For quantifying risk purposes, it is essential to have the

appropriate probability model and the standard for model selection needs to be derived accordingly.

- *Vehicle Re-identification Using Partial Trajectories in a Connected Vehicle Environment*

Connected vehicles would pack their trajectory data and send it to RSUs through Basic Safety Messages (BSM). In order to protect vehicle user's information, the vehicle ID is randomly generated every 300 seconds and the same vehicle may have totally unrelated IDs. This would result in incomplete trajectory pieces. For performance measure and trajectory prediction purposes, complete trajectories are important prerequisite and should be reconstructed if possible. The key is to re-identify vehicles with different IDs. It is possible to achieve the matching goal by investigating vehicle trajectory's spatial and temporal patterns. Candidate algorithms need to be evaluated and validated in different scenario settings. The ideal matching algorithm should be able to re-identify vehicle with high accuracy in most situations to ensure the success of reconstructing complete trajectories.

- *Short-term Vehicle Trajectory Prediction in a Connected Vehicle Environment*

The Basic Safety Messages (BSM) as well as the Signal Phase and Timing (SPaT) data contain both the vehicle (GPS positions, speed, etc.) and signal information. These data are valuable resources for improving vehicle safety performance. Since the human reaction to emergent situation is around 1.5 second, short-term vehicle prediction could help identify potential conflict trajectories before people could realize the dangerous. The prediction method should take advantage of vehicle's history trajectory and its surrounding environment information, including

nearby vehicle's dynamics and each phase's signal status. The prediction accuracy for such a task is demanding as it is safety-critical and closely related to potential crash accidents.

1.3 Dissertation Organization

The remaining chapters of the dissertation are organized as follows:

Chapter 2 is a literature review part about crash accident analysis and connected vehicle trajectory research. Modeling approaches using crash data for severity and frequency analysis will be reviewed. Related critical factor identification methods will be presented. Past research work on travel time reliability and crash risk will also be examined. Research work using vehicle trajectory data for vehicle re-identification and future trajectory prediction will be reviewed at last.

Chapter 3 is about the research of crash accident severity analysis, including crash data characteristic identification method, critical factor identification method and counter-measure safety strategy discussion. First, the method of using Mosaic plot and related statistical test will be introduced for differentiating crash injury data. Then, the proposed method for critical factor identification and quantification will be presented and validated. Finally, by applying the proposed method to 6 years cumulated crash data, a case study is conduced and corresponding safety measures are come up and discussed.

Chapter 4 will describe the research work on crash frequency analysis, including spatial temporal model for travel time imputation, data integration framework, critical corridor segment identification method and quantitative crash risk assessment method. First, the Dynamic Linear Model will be presented and validated for imputing incomplete travel time. Based on this, travel time reliability measure buffer time is derived. Together with weather information, modern

variable selection will be introduced for critical segment and weather station identification. A mixed Poisson regression family is also presented and a corresponding distribution selection method is described for crash risk assessment. At last, a corridor-level crash prediction method is proposed and validated by real crash data.

Chapter 5 will focus on the research of vehicle re-identification problem. First, a general vehicle re-identification scenario in a connected vehicle environment is introduced. Then, several machine learning algorithms are implemented to trajectory classification. A designed experiment is conducted considering various combinations of other factors besides the machine learning algorithm itself. The Bayesian framework for experiment analysis and model validation is then presented and conclusions are made accordingly. Finally, some other vehicle re-identification approaches will be discussed.

Chapter 6 will present the research work of vehicle trajectory prediction in a connected vehicle environment. First, the Kalman filtering method will be implemented as a baseline model and the parameter estimation methods will be introduced. Then, two Recurrent Neural Network based model will be presented for connected vehicle trajectory prediction. A designed experiment and corresponding error analysis will be developed for model validation.

Chapter 7 provides summaries and contributions of the dissertation and shows the potential directions for further research.

2 Literature Review

This chapter gives a comprehensive literature review of transportation safety analytics including the following topics:

1. Crash Accident Severity Analysis
   (1) Modeling Approach Overview
   (2) Critical Factor Identification
2. Crash Accident Frequency Analysis
   (1) Modelling Approach Overview
   (2) Relationship between Travel Time and Crash Risks
3. Vehicle Re-identification in a Connected Vehicle Environment
   (1) Vehicle Re-identification Using Different Technologies
   (2) Experiment Design on Non-normal and Imbalanced Datasets
4. Vehicle Trajectory Prediction in a Connected Vehicle Environment

2.1 Crash Accident Severity Analysis

2.1.1 Modeling Approach Overview

Generally, the probit model and logit model families are the two most popular methodologies to characterize crash severity. An Ordered Probit (OP) model was used to analyze injury severity by the LTCCS datasets including approximately 1000 samples ranging from 2001 to 2003 (Zhu and Srinivasan, 2011a). The large truck crash data was examined to determine the factors affecting the overall injury severity and the explanatory factors include the characteristics of the crash, vehicles and drivers. They concluded that the age, ethnicity and height are significant truck driver factors

and distracted drivers are more likely to be involved in severe crashes. Xie et al. (2009) compared Ordered Probit (OP) model and Bayesian Ordered Probit (BOP) model in exploring the relationship between crash injury severity and covariate factors using the data from National Automotive Sampling System General Estimates System (NASSGES). The dataset contains 76994 useable crash records happened in 2003. They reported that BOP had smaller prediction error than OP when the sample size was small. But when the sample size was large, BOP and OP had identical covariate coefficient estimation. However, the power of the trained model (60% prediction error as reported) is no better than a random guess due to the imbalanced structure of the training data set. The reason is that the trained model will classify all new crash records into the dominating No Injury category to achieve the smallest prediction error and therefore lacks the ability to detect other injury categories. On the other hand, the logistic model has been found to be more robust and adapts well for a wide variety of underlying assumptions about the explanatory variables (Anderson, 1972; Press and Wilson, 1978). Recently, the mixed logit model has become popular in injury severity modelling. It allows modelling the unobserved heterogeneity across individuals. The mixed logit model was used for multiple classification of truck driver injury severity considering the model differences between single and multi-vehicle accidents (Chen and Chen, 2011). In that research, the data is from HSIS database, including nearly 20000 crash records happened from 1991 to 2000 on Illinois rural highways. Direct pseudo elasticity was used for marginal effect analysis. A similar study was conducted using an Alabama truck accident database and the data included 8171 accidents happened from 2010 to 2012 (Islam et al., 2014). Both studies found some factors are critical in severe crashes, including the fatigued drivers, driver age, time of the day, weather, etc. To identify the random covariates, a simple t hypothesis test was commonly used for pre-assumed coefficient distribution (Milton et al., 2008).

2.1.2 Critical Factor Identification

Critical factor identification problem can be treated as a variable selection problem. Traditional variable selection methods include the Best Subset Selection and the Sequential Selection (the Forward Selection, the Backward Elimination and the Stepwise Selection). These methods are highly variable due to discreteness and asymptotic theories are hard to establish for making inferences (Friedman et al., 2001). Besides, the computation limit is of concerns when the number of variables is large. This limit is also known as the "Curse of Dimensionality" (Bellman, 1961). When the number of predictor variables approaches or exceeds the number of data, the Iteratively Reweighted Least Squares (IRLS) estimators for logistic and probit models have high variance and no unique solutions (Friedman et al., 2001). These situations are common in the crash related literature. Crash data is collectively expensive to obtain, and large datasets are not available. But the number of potential factors is large, including equipment, humans, roads, weathers, traffic conditions, etc. This dilemma significantly limited the model power and further rendered critical safety factors not identified.

2.2 Crash Accident Frequency Analysis

2.2.1 Modeling Approach Overview

Crash frequency research generally applies count data modeling approaches for quantitative analysis. The parametric methods are most implemented and the Poisson regression model has been widely applied in this area for decades (Joshua and Garber, 1990; Jovanis and Chang, 1986; Miaou, 1994). The classical model, however, cannot handle the data exhibiting over-dispersion and excess number of zeros. Remedies for this model come in various forms and properties. One

important Poisson regression extension is the negative binomial regression which was proposed to overcome possible overdispersion problems. Many applications of such are made in crash frequency analysis (Chang, 2005; Dong et al., 2014a; Venkataraman et al., 2011). Similarly, another extension Poisson Inverse Gaussian regression was also used and found adapted well to the situation where data is highly dispersed (Meng and Qu, 2012; Zha et al., 2016). To account for the data with a significant amount of zeros, the Zero Inflated Poisson models were developed and successfully applied in crash analysis (Dong et al., 2014b; Shankar et al., 1997). Other than the traditional modeling approaches, Neural Network and Support Vector Machine are also applied into the crash density function estimation, but the disadvantages are their inflexibility to be generalized to other data sets and their inability to provide interpretable parameters (Lord and Mannering, 2010). For the crash frequency analysis, the research goal is often to determine which factor contributes to crashes and how much it contributes. As the above introduced Poisson regression extended models all have their own regression forms, even for the same factor, different coefficients can be derived. One example is an investigation of crash frequency estimates among 3 different Poisson regression derived models (Aguero-Valverde, 2013).

2.2.2 Relationship between Travel Time and Crash Risks

It is generally known crash incidents may lead to travel time degradation and increase the buffer time (Lomax et al., 2003). A buffer time decomposition study (Kwon et al., 2011) using the data from a 30.5-mile segment of Interstate-880 highway reports that the traffic crash incidents contributed significantly to buffer time increases. In return, traffic congestion may also increase the crash probability. A study using simulated data was conducted to support such conclusion (Shefer and Rietveld, 1997). 82 cases of crash data from a 12-mile segment of Interstate-5 were

investigated in terms of impacts of traffic oscillation and the result verified this perspective (Zheng et al., 2010). However, some studies results contradicted this positive correlation. Crash data from a 16-mile segment of Interstate-94 was examined in terms of the relation with traffic congestion(Zhou and Sisiopiku, 1997). The authors concluded that more severe injury and fatal accidents tended to decrease in congested links. Another investigation of M25 London orbital motorway crash incident data claimed that traffic congestion had no impact on the frequency of accidents (Wang et al., 2009). Such inconsistent conclusions suggest that the crash risk and the degree of traffic congestion are not always positive correlated and different cases would have relationships of their own. Meanwhile, it reminds us that more data should be included and analyzed, since most above literatures only using parts of corridors' data. Noland and Quddus (2005) argued that most link specific modelling approaches were not sufficient and more research was needed to fully understand the interactions between congestion and road safety.

2.3 Vehicle Re-identification in a Connected Vehicle Environment

2.3.1 Vehicle Re-identification Using Different Technologies

Significant research has been focused on maintaining the vehicle ID as it travels through multiple video camera sites. Zeng and Crisman (Zeng and Crisman, 1997) proposed to match vehicles using body colors and the matching accuracy was 16.42% with an 8-vehicle test sample. Kogut and Trivedi (Kogut and Trivedi, 2001) investigated color features and the spatial organization of vehicle platoons together for vehicle matching. They achieved an accuracy rate of 45%. But their research work was based on a small number of samples and did not consider the site location impact on matching accuracy. Later, additional vehicle features were considered (MacCarley, 2001), such as external dimensions, points of optical demarcation, etc. Another approach was

based on a sparse representation algorithm which was originally proposed as a facial recognition method (Wang et al., 2012). The authors claimed an accuracy of 57.84%. The limitations of video-based vehicle matching methods include illumination variations, occlusion phenomena and object overlapping. Such difficulties, and the cost of video detector deployment and maintenance, make this vehicle matching method unfavorable for transportation engineering practitioners. Some research has used unique vehicle IDs for identification. Examples include matching unique Bluetooth MAC addresses (Barceló et al., 2010; Brennan Jr et al., 2010; Haghani et al., 2010; Hainen et al., 2011; Li and Souleyrette, 2016; Quayle et al., 2010; Richardson et al., 2011), GPS device addresses (Hofleitner et al., 2012) and magnetic signatures (Charbonnier et al., 2012; Kavaler et al., 2011; Kwong et al., 2009; Sanchez et al., 2011). However, not all these ID matching facilities are widely deployed in the U.S.. Another approach is vehicle trajectory tracking. The general routine is to detect vehicles in a picture by feature classification and then to track the object using Kalman filtering and its variants. A thorough literature review was conducted in on-road vision-based vehicle detection, tracking, and behavior understanding by Sivaraman, et. al. (Sivaraman and Trivedi, 2013). Their review found that the tracked object is always confined in one camera's view angle and only suitable for continuous vehicle trajectory estimation.

2.3.2 Experiment Design on Non-normal and Imbalanced Datasets

The traditional experiment analysis is the Analysis of Variance (ANOVA) approach. Even though this classical framework has served the simulation and field data analytics for decades, it also has two limitations. First, this approach requires well balanced sample sizes with the number of observations for each set of conditions being almost equal. In a real data collection process, observations are often missing due to human error, database malfunction, and condition

limitations. Some modifications to the ANOVA were proposed by using approximate methods (Searle and Gruber, 2016; Speed et al., 1978), but they are only appropriate for situations where the dataset is mildly unbalanced (Montgomery, 2017). Second, non-normal data is usually transformed to make it appear normal (Box and Cox, 1964). The transformation should be conducted very carefully, because it can fail if count data contains many zero values, model assumptions are violated, or the scope of inference is limited. To account for these disadvantages, the Residual Maximum Likelihood (REML) method is often employed for parameter estimation as an alternative to the Maximum Likelihood (ML) method when both random and fixed factors are considered in the General Linear Mixed Model (GLMM) (Corbeil and Searle, 1976). Since the likelihood equations are in highly non-linear, closed analytical solutions are not available. Approximation methods include the Quasi-Likelihood (QL) approach which uses a Taylor series expansion (Breslow and Clayton, 1993; Schall, 1991; Wolfinger and O'connell, 1993) and the Gauss-Hermite Quadrature (GHQ) method which uses optimal subdivisions to numerically evaluate the integrand (Pinheiro and Chao, 2006). The QL approach induces bias for inference purposes and the GHQ approaches work poorly with more than two random factors (Bolker et al., 2009). Instead of trying to solving the likelihood functions, the Markov Chain Monte Carlo (MCMC) algorithm (Gilks et al., 1996) generates random samples from the distributions of parameter values of factor effects and is free of these drawbacks.

2.4 Vehicle Trajectory Prediction in a Connected Vehicle Environment

The vehicle trajectory prediction problem can be classified into 2 groups. One is the long-term trajectory prediction and the other is the short-term prediction. Some progresses for long-term trajectory prediction are made using the data-driven approach in recent years(De Brébisson et al.,

2015; Endo et al., 2017; Pecher et al., 2016). In the 2015 KAGGLE competation, De Brebisson, Simon, etc. (De Brébisson et al., 2015) published their first-place solution in the Kaggle discovery challenge on taxi destination prediction competition. The winning model was based on a variant multi-layer perceptron (MLP) architecture. Partial trajectories data and other metadata information were used as inputs and the final destination was determined by a softmax layer followed by weighted averaging of predefined destination cluster centroids. In 2016, Pecher, Hunter and Fujimoto (Pecher et al., 2016) attempted three approaches for long-term trajectory prediction or route prediction. The approaches include the Krumm's destination prediction algorithm based on efficient routes, artificial neural networks and Markov models. The T-drive trajectory dataset consisting of GPS trajectories of taxicabs in Beijing is used. All three methods heavily relied on vehicle past trajectory history and authors concluded that using trajectory data from other vehicle can substantially improve prediction accuracy. In 2017, Endo, Nishida, Toda and Sawada (Endo et al., 2017) applied the Recurrent Neural Network (RNN) structure to the vehicle destination prediction. Specifically, their method discretized trajectories into grid features and then treated them as sequences for model inputs. The taxi service trajectory (TST) dataset and GeoLife dataset are used for validating their methods. As for the short-term prediction problem, traditionally the Kalman filter and its variants are the most widely implemented methods (Prevost et al., 2007; Zhang et al., 2017). In a Unmanned Aerial Vehicle scenario (Prevost et al., 2007), an extended Kalman filter is first used to estimate the states of a moving object detected by a Unmanned Aerial Vehicle from its measured position in space. The optimal object trajectory is then predicted from the object states and using the motion model defined for Kalman filtering. Recently, the RNN was applied for modeling a moving vehicle driver's behavior (acceleration) (Morton et al., 2017), Authors use the Recurrent Neural Network to model the car following model and predicting

vehicle acceleration in a one dimensional situation. The model took the ego-vehicle previous second's state (headway, relative speed difference, speed and acceleration) as inputs and outputted parameters of a mixture Normal distribution. By sampling from the distribution, future acceleration was derived and further used to calculate vehicle speed and positions. Their results revealed that the good performance of the neural model is due to its ability to identify recent trends in the vehicle's state.

3 Crash Accident Severity Analysis

3.1 Introduction

The Interstate 10 (I-10) is one of the national most important corridor for freight transportation. By 2035 the projected the projected average daily traffic will exceed 85,000 including over 20,000 trucks (USDOT, 2007). Truck drivers are considered likely to bear higher risks of suffering serious injuries and fatalities than passenger vehicle drivers (USDOT, 2005). For the case of I-10 Arizona segment, in the past 6 years (2010-1015), nearly 25% of injured people are truck passengers or drivers according to Arizona Department of Transportation (ADOT) crash accident records. The National Institute for Occupational Safety and Health (NIOSH) (WALLER, 2003) reported that American truck drivers have as much as 7 times of greater risks of fatality and 2.5 times of greater risks of suffering an injury than average workers (Chen and Chen, 2011; Saltzman and Belzer, 2007). Studying the national freight corridor safety conditions and analyzing the causes for truck crash accidents are essential. Investigating the crash severity is traditionally considered one of most effective ways to gain insights for truck safety research. Deriving strategies for prevention and mitigating crash severity is desired for freight transportation practitioners, administrators, and researchers.

Section 3.2 gives a detailed description about the dataset used in this research topic. Section 3.3 provides the developed methodology framework for exploring accident severity. Section 3.4 provides the model validation, identified critical factors and counter-measures. Section 3.5 summaries this chapter.

3.2 Data Overview

Available Interstate 10 (Arizona Segment) vehicle crash accident data is collected from 2010 to 2015. Specifically, there are totally 25189 truck driver and passenger injury records, taking up 25% of the total injury records. The injury records are labeled by 5 levels: No Injury, Possible Injury, Non-Incapacitating Injury, Incapacitating Injury and Fatality. For truck related records and non-truck related records, the injury distributions are different. This can be shown directly from Figure 3.1(a). This figure shows the proportion of injury cases in each vehicle records. The height of each injury category rectangle represents its percentage proportion in either the truck records or the non-truck records. The width of each vehicle category represents the number of its injury records. Both vehicle group have unbalanced injury level distributions: The No Injury cases dominate the record and the other injury level cases are overwhelmed in general. Meanwhile the size of the non-truck injury cases is much bigger than that of the truck injury cases. To further strictly test how the injury distribution is related to the vehicle type, a Chi-square test is conducted. The null hypothesis of the test is that the vehicle type and the injury level are independent variables. Let n_{ij} be the observed number of injury cases of i th vehicle group ($i=1,2$) and j th injury level ($j=1,\cdots,5$). Let $n_{i+}=\sum_j n_{ij}$ represent total number of injury cases for i th vehicle group, $n_{j+}=\sum_i n_{ij}$ represent total number of injury cases for j th injury level and $n_{++}=\sum_i \sum_j n_{ij}$ is the grand total number of injury cases. Under the null hypothesis, the expected number of injury for each scenario is $\hat{m}_{ij}=n_{i+}n_{j+}/n_{++}$. The Pearson standardized residual is $d_{ij}=\left(n_{ij}-\hat{m}_{ij}\right)/\sqrt{\hat{m}_{ij}}$ and the test statistic is $\chi^2=\sum_i \sum_j d_{ij}^2=358.66$. The test is conducted by comparing the statistic to the critical value of $\chi^2_{0.05}$ with degree of freedom $(2-1)(5-1)=4$, which is 9.49 and far smaller than 358.66. Thus we reject the null hypothesis and conclude the injury distribution is dependent on the

vehicle type. Based on Figure 3.1(a), Pearson residuals of each injury rectangle are represented by different colors in Figure 3.1(b). This figure is known as a Mosaic plot (Friendly, 1994) and is highlighting the over-represented and under-represented injury level in both truck and non-truck groups. It shows that more No Injury cases and Fatality cases appeared and less Possible Injury and Non-Incapacitating cases appeared in the truck group than the expected. The truck accidents' unique characteristic requires further investigation of factors contributing to different injury severity levels.

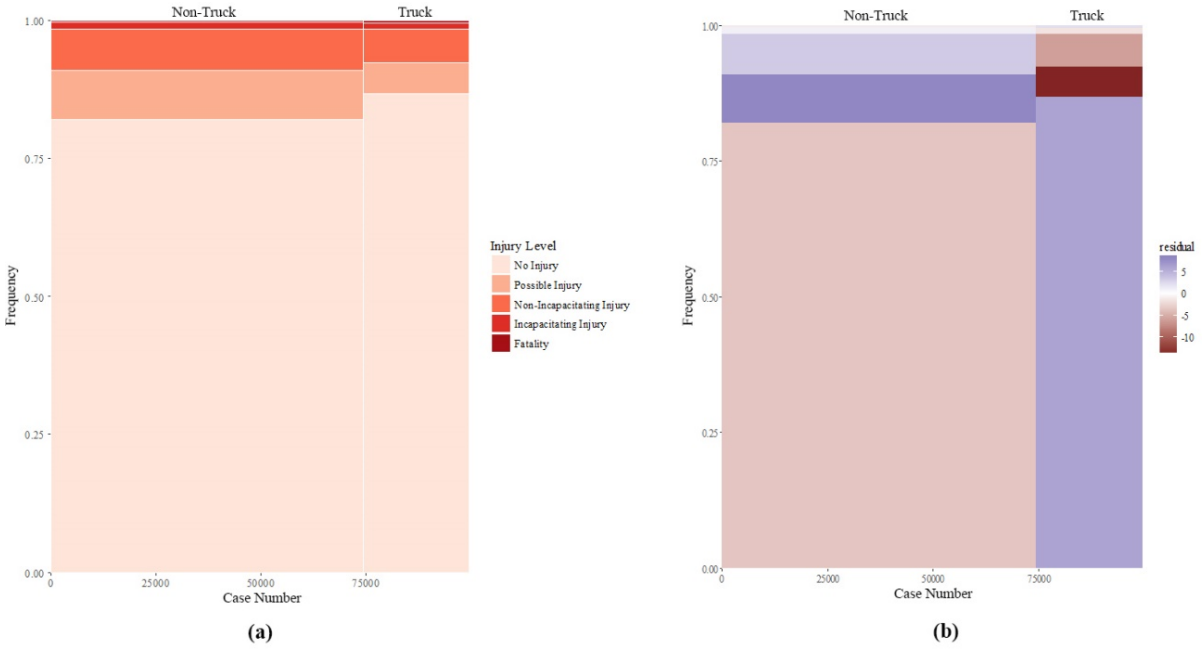


Figure 3. 1 Mosaic Plot of Injury VS Vehicle Type

The 25189 raw records cannot be fully utilized due to some factor variable values missing. After removing records with missing factors, the clean data includes 5940 No Injury cases, 436 Possible Injury cases, 288 Non-Incapacitating Injury cases and 22 Incapacitating Injury cases. Further, the

Non-Incapacitating Injury category and Incapacitating Injury category are merged into one category Severe Injury, since the Incapacitating Injury size is too small for any effective model to be trained.

There are 28 crash related categorical variables and each of them has corresponding subcategory factors. Table 3.1 shows the variable names and their sizes. The details of subcategory factors can be found in Appendix A.

Table 3. 1 Injury Related Variables and Sizes

Variable	*Size*	*Variable*	*Size*	*Variable*	*Size*	*Variable*	*Size*
Age	7	Unit Action	14	Control Type	8	Light Condition	6
Sex	2	Body Style	27	Surface Condition	7	County ID	5
Safety Device	7	Make	33	Incident Year	6	Latitude	824
Injury Status	4	Color	29	Incident Month	12	Longitude	1223
Violation	15	Estimated Speed	7	Incident Day of Week	7	Traffic Way Type	5
Road Alignment	3	Speed Over Limit	2	Incident Hour	24	Intersection Type	6
Road Grade	5	Damage Area	12	Collision Manner	9	Weather	9

3.3 Methodology

The total number of subcategory factors are large (2315 indicator functions equivalently) and identifying the most critical ones is really challenging given many of them are noise and not irrelevant. In this section, modern variable selection methods are utilized along with multinomial logistic regression to facilitate this process. The concept of relative risk is derived to interpret the resulted classifier models. In order to deal with the imbalanced data structure, a resampling technique Sub-Bagging is proposed to reach the final conclusion.

3.3.1 Feature Selection

Regularized Logistic Regression

To identify critical factors that contribute to sever injury crashes, the multinomial logistic regression classification method is used. By differentiating injury severity levels, factors will be assigned different coefficients for different injury severity. According to various combinations of these factors, the probability of a truck passenger or driver ended in severity level j (assuming K levels in total) is determined by:

$$p_j(\mathbf{x}) = \Pr(Y = j \mid \mathbf{x}) = \frac{e^{\beta_{0j} + \mathbf{x}\boldsymbol{\beta}_\mathbf{j}}}{\sum_{k=1}^{K} e^{\beta_{0k} + \mathbf{x}\boldsymbol{\beta}_\mathbf{k}}} \tag{3.1}$$

In which, $\mathbf{x} = (x_1, x_2, \cdots x_p)$ and $\boldsymbol{\beta}_\mathbf{j} = (\beta_{1j}, \beta_{2j}, \cdots \beta_{pj})^T$ are vectors representing p factors involved in the crash and their corresponding coefficients. β_{0j} is the intercept term. Given n crash records, traditionally, the coefficients are estimated by minimizing the negative joint conditional likelihood:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} -\sum_{i=1}^{n} \log p_{y_i}(\boldsymbol{\beta}_j; \mathbf{x}) \tag{3.2}$$

In which $y_i \in \{1, 2, \cdots K\}$ is the injury severity level for ith single crash record.

Given thousands of factors to consider, variable selection is necessary to ensure a sparse result only keeping critical factors. By using modern variable selection method, injury severity classification can be conducted with more parsimonious and interpretable models. The most important modern variable selection method is the known LASSO (Tibshirani, 1996). After that, efforts were made to improve this framework including the Adaptive LASSO (Zou, 2006), SCAD (Fan and Li, 2001) and Nonnegative Garotte (Breiman, 1995). Based on the LASSO, the Elastic Net (EN) was proposed (Zou and Hastie, 2005) and later the Adaptive Elastic Net (AEN) was proposed to obtain the oracle properties of estimators (Zou and Zhang, 2009). In this work, the EN and the AEN are utilized along with the logistic regression in injury severity classification. These two approaches are both belonging to the penalized logistic regression family. One big advantage of the modern variable selection methods is that complete variable coefficient paths are derived in a continuous fashion which successfully avoids the high variance brought by discreetness in classic sequential variable selection methods. That is, the factor coefficient estimation and variable selection are emerged in one procedure. The EN is actually combining the LASSO and the Ridge Regression together in a way such achieving the desired sparse selection results and meanwhile overcome some known disadvantages of the LASSO. The prominent difference between the traditional logistic regression and the improved logistic regression is how they estimate the coefficients with the previous problem settings unchanged (Equation (3.2)). Specifically, the EN estimates factor coefficients by imposing a penalty term on the object equation:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} -\sum_{i=1}^{n} \log p_{y_i}\left(\boldsymbol{\beta}_{y_i}; \mathbf{x}\right) + \lambda \sum_{j=1}^{K} P^{EN}{}_{\alpha}\left(\boldsymbol{\beta}_j\right) \tag{3.3}$$

In which the penalty term is:

$$P^{EN}{}_{\alpha}\left(\boldsymbol{\beta}_j\right)=(1-\alpha)\left\|\boldsymbol{\beta}_j\right\|^2+\alpha\left\|\boldsymbol{\beta}_j\right\|_1=\sum_{s=1}^{p}\left[(1-\alpha)\beta_{sj}{}^2+\alpha\left|\beta_{sj}\right|\right] \tag{3.4}$$

Tuning parameters λ and α are determined by a grid search with 5-fold Cross Validation. The purpose of the imposed penalty terms is to reduce the complexity of derived logistic regression model by soft thresholding of β_{sj} s among which small coefficients are shrunk to zero to achieve variable selection effects. For the AEN, the adjustment is made on the penalty term, which is assigned with different weights to the l_1 term:

$$P^{AEN}{}_{\alpha}\left(\boldsymbol{\beta}_j\right)=(1-\alpha)\left\|\boldsymbol{\beta}_j\right\|^2+\alpha\left\|W_j\boldsymbol{\beta}_j\right\|_1=\sum_{s=1}^{p}\left[(1-\alpha)\beta_{js}{}^2+\alpha w_{js}\left|\beta_{js}\right|\right] \tag{3.5}$$

In which $w_{js}=\left|\hat{\beta}_{js}^{enet}\right|^{-\gamma}$. $\hat{\beta}_{js}^{enet}$ is the estimator of factor coefficients by the EN logistic regression and γ is a constant assigned 1 in this work. The advantage of the AEN is to allow prior knowledges to be incorporated in variable selection or to explore the influence of some specific interesting factors for which weights are assigned 0 and no penalty is exerted in the variable selection process.

Fitting Process

The model fitting is conducted by R package (Friedman et al., 2016; R Core Team, 2016) and detailed computation procedures should be referred to Friedman et al. (2010). The EN and AEN regularized logistic regressions are conducted respectively. For each model, variables are selected through coefficient shrinkage. Take the EN case for example, Figure 3.2 shows the variable coefficient selecting process. The number of effective variables decreases as the tuning parameter λ increases with a fixed α. The choices of λ and α are done based on a grid search method with a 5-fold cross validation using the training data set. In this set of models, the numbers of selected factors are 298 for the No Injury level, 281 for the Possible Injury level and 242 for the Severe Injury level, while the original number of indicator variables is 2300 for each injury level. Nevertheless, aggregating 1000 replicates of such experiment and using proper interpretation are necessary before any comprehensive conclusion can be drawn.

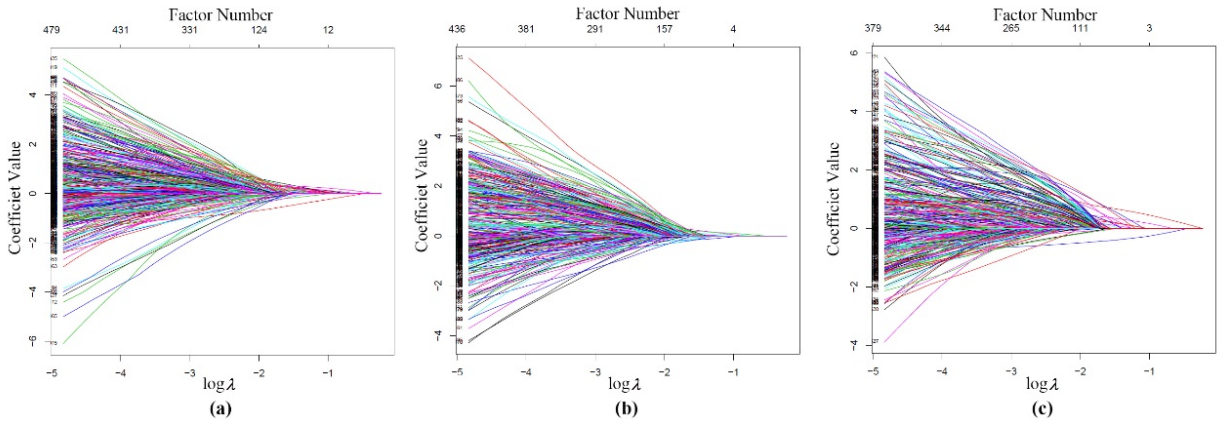


Figure 3. 2 Coefficient Shrinkage

Model Interpretation

The Elasticity (for continuous variables) and the Direct Pseudo Elasticity (for categorical variables) analysis were used for model interpretation with regard to the effects of specific variables on outcome probabilities (Chang and Mannering, 1999; Shankar and Mannering, 1996; Ulfarsson and Mannering, 2004). However, the Elasticity analysis cannot differentiate the effects on multiple class outcomes directly. For example, a factor's elasticity may increase or decrease both Possible Injury and Severe Injury probabilities at the same time. This makes it difficult to differentiate the factor's roles across injury levels without some ad hoc analysis as remedies. An typical example might be an airbag deployment which is addressed by Ulfarsson and Mannering (2004). Thus instead, the concept of relative risk is proposed in this research to overcome such drawbacks. According to the parametrization in Equation (3.1), the log odds of two different injury levels is:

$$Log\frac{\Pr(Y=j\,|\,\mathbf{x})}{\Pr(Y=K\,|\,\mathbf{x})}=\beta_{0j}-\beta_{0K}+\mathbf{x}(\boldsymbol{\beta}_{\mathbf{j}}-\boldsymbol{\beta}_{\mathbf{K}}) \tag{3.6}$$

By exponentiating the log odds, the ratio of two probabilities is interpreted as the relative risk. Take one EN regularized logistic regression models for instance, in the estimated probability function $p_{Severe\_Injury}(x)$, the coefficient of the indicator variable indicating whether the estimated speed is over 80 miles per hour is 4.26. Correspondingly, the coefficients of the same factor in $p_{No\_Injury}(x)$ and $p_{Possible\_Injury}(x)$ are -1.29 and 0.57. Then, given the remaining factors unchanged, an increase in both relative risks is calculated as below given the speed is changed to over 80 mph from below 30 mph (the reference level):

$$\frac{\Pr\left(Y = SevereInjury \mid \mathbf{x}, x_{I(speed>80mph)}\right)}{\Pr\left(Y = NoInjury \mid \mathbf{x}, x_{I(speed>80mph)}\right)} = e^{4.26+1.29} = 257.24 \tag{3.7}$$

$$\frac{\Pr\left(Y = SevereInjury \mid \mathbf{x}, x_{I(speed>80mph)}\right)}{\Pr\left(Y = PossibleInjury \mid \mathbf{x}, x_{I(speed>80mph)}\right)} = e^{4.26-0.57} = 40.04 \tag{3.8}$$

The passengers and drivers in a truck will have a relative risk of severe injury vs. no injury increased by nearly 257 times and a relative risk of severe injury vs. possible injury increased by nearly 40 times. Similar interpretation procedures can be applied to all crash factors including both the selected and the unselected. For the unselected, the coefficients are 0s. Based on this interpretation, crash related factors are divided as contributing factors and alleviating factors. The contributing factors are those which both increase the relative risks of Severe Injury vs. No Injury and Severe Injury vs. Possible Injury by more than 1 time and the alleviating factors are those which both decrease the two relative risks by more than 1 time. For example, in one fitted model, the contributing factors and alleviating factors are identified in Table 3.2. But critical factors are not identified merely according to Table 3.2, instead final conclusions are drawn with help of the Bagging technique in the next section.

Table 3. 2 Contributing and Alleviating Factors by 1 Set of Trained Models

Contributing factors	*Alleviating factors*
Speed>80mph	Collision Manner Sideswipe Same Deriction
70mph<Speed<80mph	Shoulder And Lap Belt
Damage Area 9	Truck Brand 1
Damage Area TOTALED	Truck Body Style TK
Surface Condition Wet	
Surface Condition Snow	
Truck Brand 2	
Truck Brand 3	
Air Bag Deployed	

3.3.2 Sub-Bagging

The crash record data structure is extremely imbalanced. The dominating No Injury class takes up to nearly 90% of the total complete sample data. Using all of them together with the other 2 classes of injury observations will cause serious masking problem. That is, the trained model has little power to detect the minor classes, since classifying all observations to one dominating class will achieve the minimal prediction error. However, for our purpose, the trained model's power to differentiate injury levels is essential because the resulted models should provide information about how factors are affecting severe injuries. Traditionally, there are two ways to deal with the imbalanced data structure. One is called case control subsampling technique by which a logistic regression model is fitted on subsampled data and this method only needs a simple adjustment to its resulted model intercept term (Anderson, 1972). Another way to improve the standard case control subsampling is to weight the subsampled observations by the inverse of their probability of being sampled (Horvitz and Thompson, 1952). However, the first technique has much bias in estimate and the second one introduces more variance (Fithian and Hastie, 2014). Both of them did not make full use of the data and possibly information is lost in subsampling, therefore introducing bias and variance.

The promising Bagging (Bootstrap Aggregation) approach serves to improve unstable procedures (Breiman, 1996). The idea is to generate bootstrap samples, obtain many base estimates and make predictions based on all classifiers (Friedman et al., 2001). Here, we propose to combine it with the case control subsampling technique and name it as Sub-Bagging. Table 3 shows the whole procedure of this Sub-Bagging algorithm. First, the data set is divided into 2 parts, 2/3 as the training set for model training and 1/3 as the test set for model validation purpose. Let the training data set be $\mathbf{Z} = \{(\mathbf{x_1}, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ and n_1, n_2, n_3 be each category size. The No Injury class bootstrap samples are generated according to an acceptance probability $1/n_1$. The newly generated training data set is $\mathbf{Z}^b = \{(\mathbf{x_1}^b, y_1^b), \cdots, (\mathbf{x}_m{}^b, y_m{}^b)\}$ with sample size $m = n_1^* + n_2 + n_3$ and n_1^* is bootstrap sample size for the dominating crash injury category. This is a hyperparameter which needs to be tuned in fitting model. The superscript $b = 1, 2, ..., B$ represents the order of sampling process. In this research, the available crash record data size for each injury category is 5940, 436 and 288. After taking 1/3 as test data for validation purpose, we set $n_1^* = n_2 = 290, n_3 = 192$. Such sample sizes make the data more balanced then the original one. For each of the generated training data set $\mathbf{Z}^b$, logistic regression models are fitted with variable selection methods. The corresponding classifier is:

$$\hat{f}^b(\mathbf{x}) = \arg\max_j \hat{Pr}^b(Y = j \mid \mathbf{X}) \tag{3.9}$$

Totally $B = 1000$ bootstrap classifiers are generated, then the Bagging classifier is summarizing all of them and making classification according to the number of 'votes' of each category received from 1000 bootstrap classifiers. At the same time, contributing and alleviating factors given by each of these classifications are pooled together. By counting times of each factor's appearances, the final critical factors are identified. The more times one factor appears as a contributing factor or as an alleviating factor, the more important it is. The advantage of employing this technique here is making sure the dominating category records are fully used without information loss and meanwhile keeping the trained model's power to detect all crash injury categories.

Table 3. 3 the Sub-Bagging Algorithm

Algorithm *Sub-Bagging*

(1) Split the data into the training set and the test set

(2) Repeat for b=1,2,…,1000

i) Generate the bootstrap sample $\mathbf{Z}^b$ from the training dataset with sample sizes n_1^*, n_2, n_3 for each injury level

ii) Perform the variable selection method using Equation (3-5)

iii) Perform the relative risk analysis for each factor using Equation (6)

iv) Validate the generated model (Equation (1)) by predicting injury levels in the test dataset

(3) Aggregating models generated in Step (2)

i) Summarizing the critical features by count

ii) Validate the aggregated modeld (Equation (9)) by predicting injury levels in the test dataset

3.4 Results

In this section, results of model analysis are presented. First, different modern variable selection methods are validated and compared in terms of their prediction errors. Then, based on the EN model, critical factors are identified. Further, effects of factors of particular interest are quantified with the relative risk interpretation. Last, safety-improvement strategies are discussed with regard to those identified safety critical factors.

3.4.1 Model Validation and Comparison

Two kinds of variable selection method (the EN and the AEN) were introduced in previous sections. In this research, the AEN weights for l_1 norm of coefficients are set as $w_{js} = \left|\hat{\beta}_{js}^{enet}\right|^{-\gamma}$ in Equation (3.5). Meanwhile, to explore the effects of high speed factor and over limit speed factor on truck crash injury, another different set of weights are created with no penalties on indicator variables representing whether the truck's estimated speed was over 70 mph and whether the truck's speed was over limit. The remaining weights are still kept unchanged as in the EN approach. For simplicity, this new variable selection approach is denoted as EN1 to indicate its slight difference from the EN. Further, the 3 variable selection methods and their Sub-Bagging improved models are compared using the test data set in terms of their overall prediction errors and category specific prediction errors. The definition of prediction error is:

$$\text{Prediction Error} = \frac{\text{number of incorrected predictions}}{\text{total number of test observations}} \tag{3.10}$$

The boxplots in Figure 3.3 shows the comparison of 3 logistic regression models with different variable selection mechanisms. According to their prediction error medians (solid black lines crossing boxes), the EN and the EN1 are more accurate than the AEN in the overall performance, while the AEN dominates in detecting Possible Injury class. The EN and the EN1 perform best in detecting No Injury class and Severe Injury class, respectively.

By using the Sub-Bagging technique, the 3 methods are improved in prediction accuracy. This is shown in the Figure 3.4's bar plots. Medians of prediction errors of original the AEN, EN and EN1 models are used in comparison with the Sub-Bagging models. All 3 variable selection methods have improved with 10.1%, 4.4% and 12.2% decrease respectively in overall prediction errors. Particularly, with the Sub-Bagging technique the EN and the EN1 improved logistic regression model can achieve an overall prediction error as small as 0.365 and 0.386. Specifically, the Sub-Bagging helps decrease the prediction errors in most injury levels, though there are two exceptions for the No Injury class (Figure 3.4 (1) and (3)).

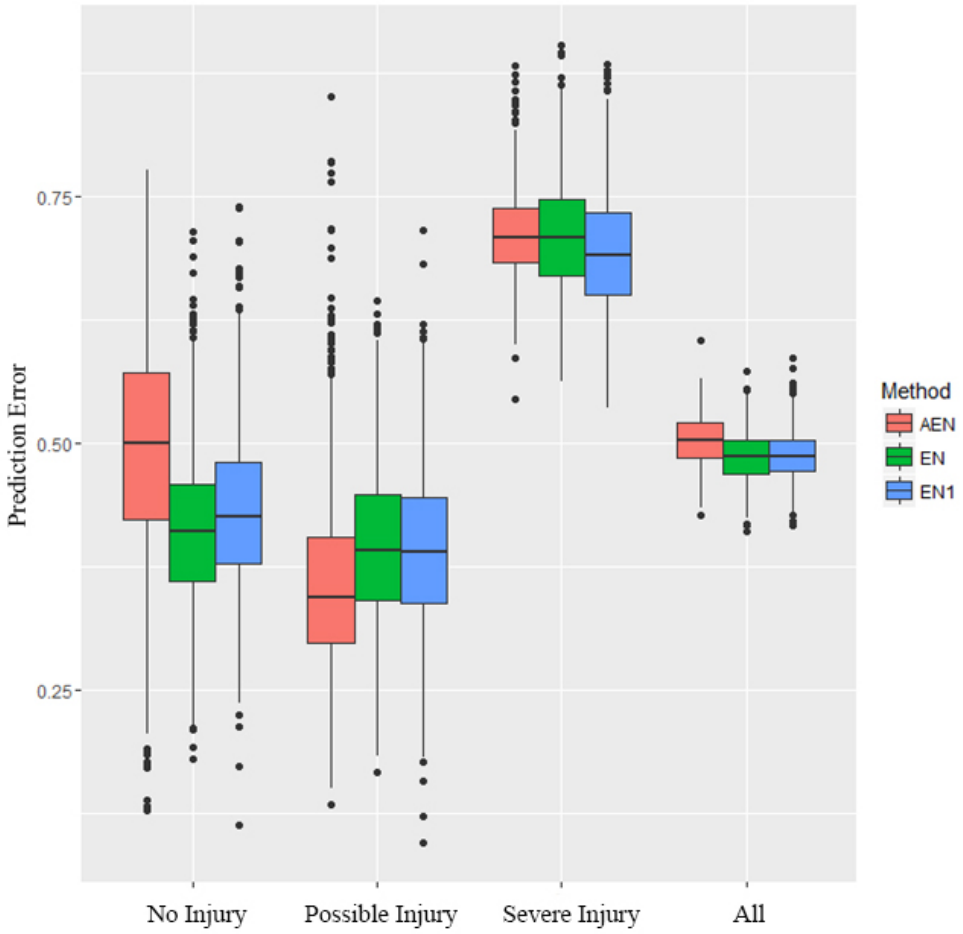


Figure 3. 3 Boxplot of 1000 Classifiers' Prediction Error Comparison

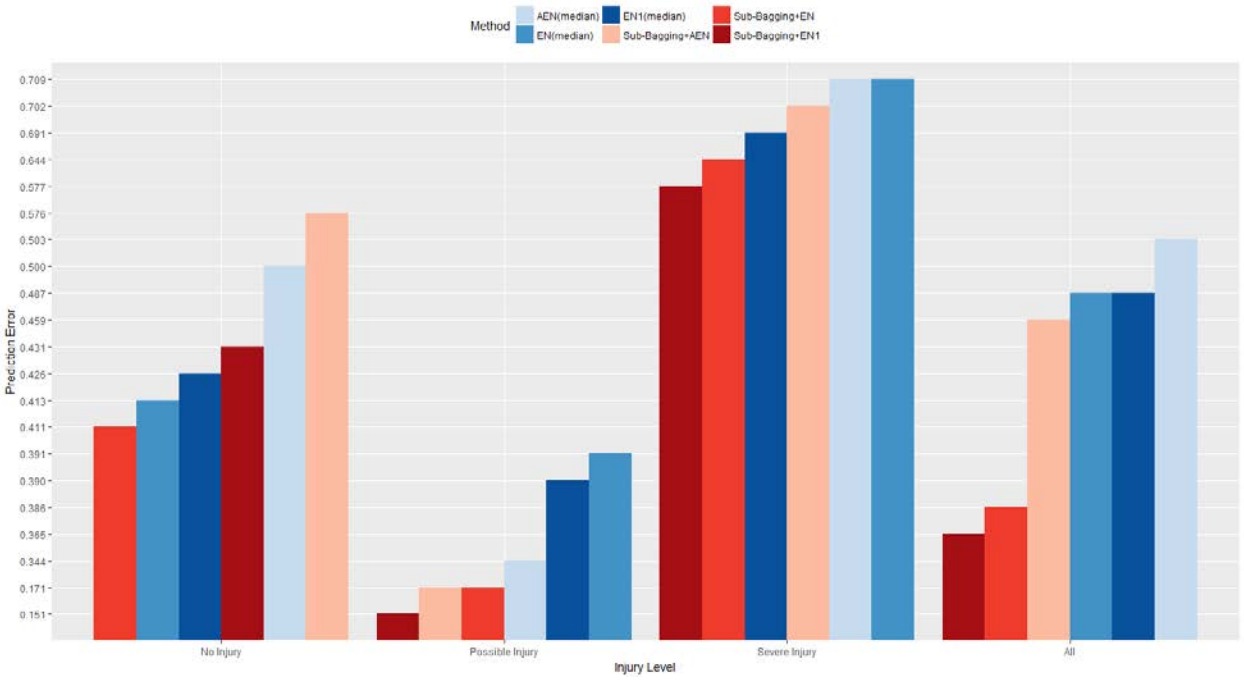


Figure 3. 4 Bar Plots of Prediction Error Comparisons

3.4.2 Critical Factor Identification

Critical Factor Summary

According to Figure 3.4, by using the Sub-Bagging, the EN (the red bar) and the EN1 (the dark red bar) will result better overall prediction accuracies than the AEN (the pink bar). The EN1 variable selection method is based on the EN and our own preference of interest and its prediction error (with the Sub-Bagging) is only slightly smaller than the EN (with the Sub-Bagging). Therefore, in order to avoid losing generality of this algorithm, the following critical factors are identified based on the EN variable selection with the Bagging technique and the model users can always incorporate their own preference of interest by revising the weights for the l_1 norm of factor coefficients in Equation (3.5). According to the Sub-Bagging algorithm (Table 3.3), the more frequent one factor appears as a contributing factor or an alleviating factor, the more

important it is. Since the longitude and the latitude factors are easier to see in a map, the GPS factors (the longitudes and the latitudes) and the Non-GPS factors are presented in different ways.

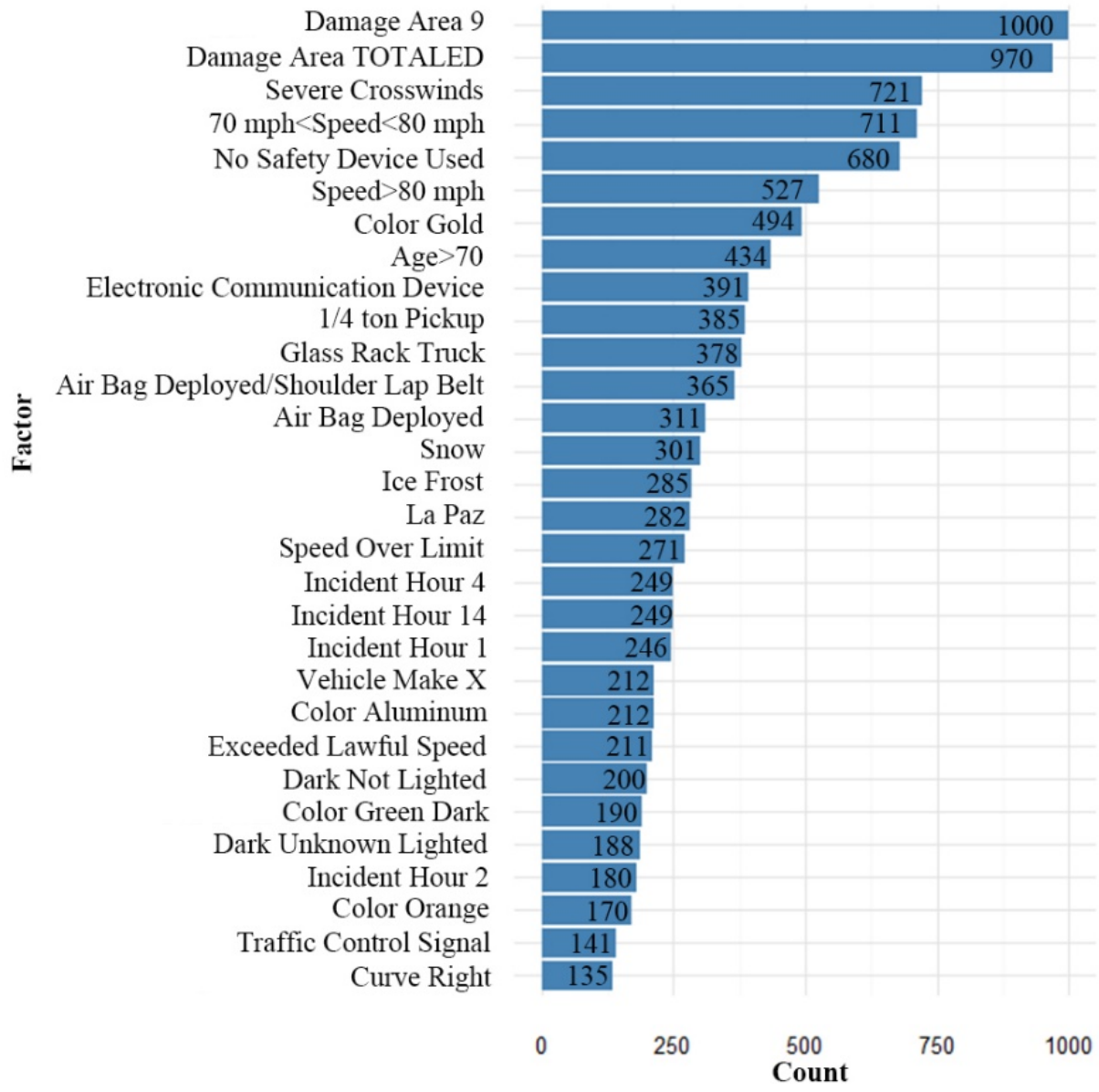


Figure 3. 5 Identified Non GPS Contributing Factors

Figure 3.5 displays the top 30 most counted Non-GPS contributing factors. As for the alleviating factors, the number of identified factors by each logistic regression model is far less than that of identified contributing factors. Figure 3.6 displays the top 10 most counted Non-GPS alleviating factors.

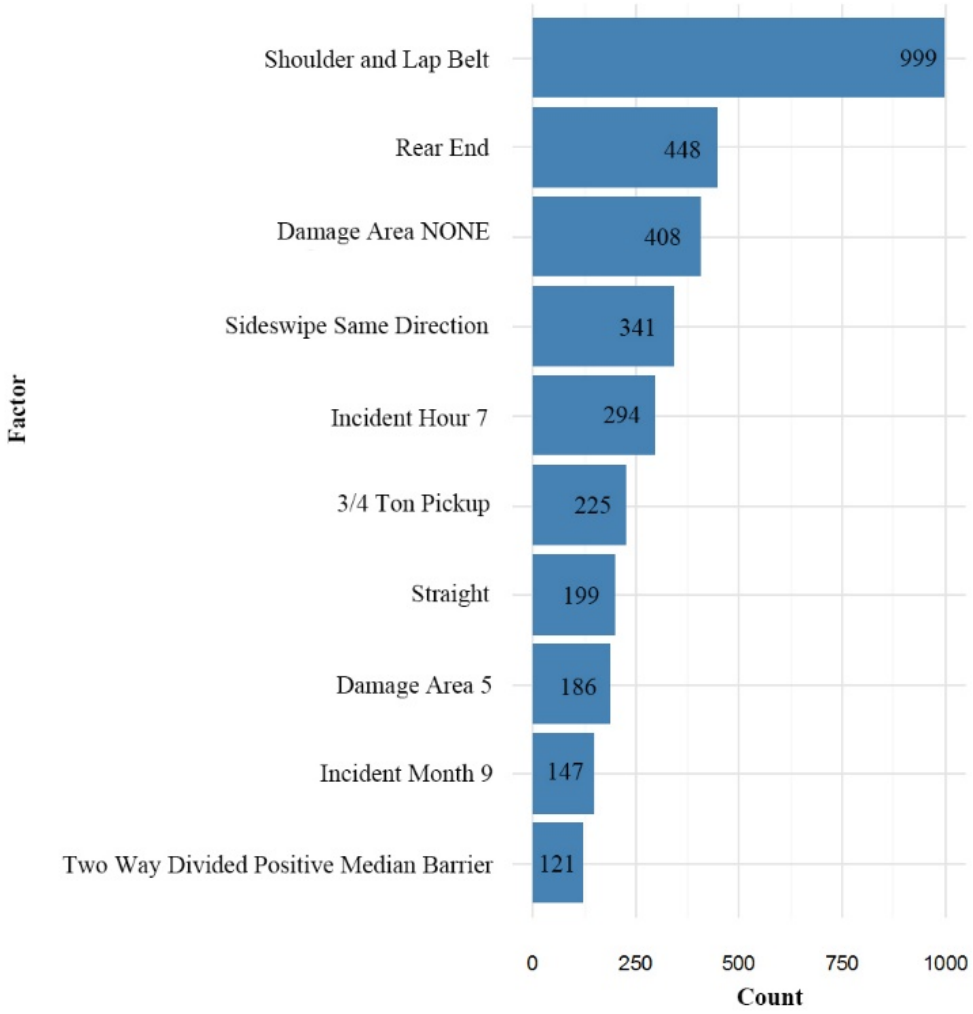


Figure 3. 6 Identified Non GPS Alleviating Factors

Fig 3.7 displays a contour plot of count of high risk GPS locations which are highlighted. The location coordinate consists of a latitude and a longitude. The contour shows the latitudes and the

longitudes appearing many times as contributing factors. The black curve is the shape of I-10 across Arizona and the road segments in the highlighted regions are considered highly risky. There are few identified alleviating GPS locations, so there is little variation if they are plotted in a contour which is omitted here.

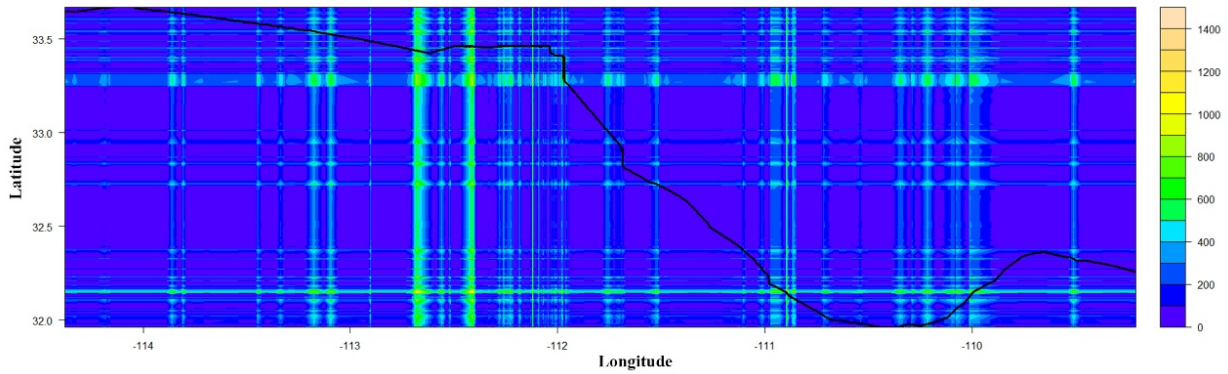


Figure 3. 7 Identified GPS Contributing Factors

Quantification of Factors of Interest

The AEN provides a framework in which the model users can make their preferred factors of interest not shrunk at all in the variable selection process and therefore to better observe and explore their quantified effects on crash injury severity. 0 weights are assigned to the preferred factors to ensure no penalty is put on them. In our experiment, the EN1 is such a variable selection process. 3 factors are chosen as the interested factors. For every logistic regression model using bootstrap samples in the Bagging procedure, the relative risks of Severe Injury vs. No Injury and Severe Injury vs. Possible Injury can be derived. The median of each group of relative risk is used to measure the effects of the related crash factors.

Table 3.4 Quantified Effects of Preferred Factors

factors of Interest	*Relative Risk*	
	Severe Injury VS Possible Injury	*Severe Injury VS No Injury*
70mph<Speed<80mph	2.28	2.92
Speed>80mph	3.78	36.75
Speed Over Limit	0.95	1.18

Table 3.4 displays the quantified results. It is found that if a truck is running at a speed over 80 mph, the chances for the driver and passengers to suffer severe injuries will greatly increase to the degree that the relative risks of Severe Injury vs. No Injury and Severe Injury vs. Possible will gain 36.75 times and 3.78 times respectively comparing with a truck running under 30mph (the reference level). Meanwhile the speed between 70 mph and 80 mph will also increase both relative risks more than 2 times. And the behavior of speeding will increase both risks by nearly 1 time. This quantification procedure can also be applied to other users' preferred factors in practice.

3.4.3 Analysis and Strategies for Safety Improvement

Figure 3.5, 3.6 and 3.7 present the critical factors having significant influence on truck crash injuries. Based on the results, some safety initiatives can be come up aiming at reducing safety risk. The vehicle color, make and brand are mostly decided by people personal preference and therefore the related identified information is useful for insurance companies as references.

The damage areas of a crash involved truck is really important. "Damage Area 9" and "TOTALED" are the top 2 contributing factors for severe injuries. Meanwhile Fig. 3.6 reveals that if the damage is occurred to "Area 5" and "NONE", the possibility for the severe injury to happen will be decreased. According to the Arizona Crash Report Forms Instruction Manual (ADOT, 2017), "Damage Area 9" and "Area 5" are corresponding to the vehicle's roof and rear part

respectively. This information is useful for truck manufacturers helping them make safer and more reliable trucks.

The identified contributing speed factors include “speed over 70 mph” and “speed exceeding limits”. These factors will increase the chance of truck drivers and passengers to be severely injured. To efficiently control the speed, one promising technology is called speed harmonization which provides the driver with speed recommendation based on traffic conditions and weather information(Chira-Chavala and Yoo, 1994; Talebpour et al., 2013).

The weathers of “Severe Crosswinds” and “Snow” are two factors identified contributing to severe crash injuries. These risky weathers could cause potential property and life loss on the I-10 highway. Thus, alerts for such weather forecasting should be distributed to truck drivers via radio channels or mobile phone messages in time. Once the drivers are notified, they can reschedule their freight delivery plan. A roadway surface condition factor “Ice Frost” is also closely related to the snow weather. Thus, timely defrost the ice and frost or temporarily close the ice frost covered roadway segment are necessary to improve freight corridor safety.

The identified contributing incident hour factors are 1 am, 2 am, 4 am and 2 pm while the identified alleviating incident hour factor is 7 am. This is indicating that trucks operated in the night are very likely to incur severe injury crash accidents. Particularly, driving in the late night will probably make drivers tired and distracted. Therefore, freight truck drivers should make their operation time during the daytime and avoid fatigue driving late in the night if possible.

Safety devices factor “No Safety Device Was Used” and “Air Bag Deployed” are identified as contributing factors and “Shoulder and Lap Belt” is identified as an alleviating factor. Normally, air bags are deployed in serious collisions and can greatly reduce the likelihood of fatal injuries.

Meanwhile, researchers found that it can cause minor abrasions and elevate injury levels(Ulfarsson and Mannering, 2004). Since there are no fatal injury training samples in this research, the factor Air Bag Deployed was only shown to have effects of elevating injury levels. Therefore, the corresponding safety strategies should be making using the “Shoulder and Lap Belt” as mandatory requirements for both truck drivers and passengers. At the same time, the quality of air bags and the way it works may need more attention for truck industry and manufacturers.

Poor lighting conditions are also identified as contributing factors for severe crash injuries. The factors “Dark Not Lighted” and “Dark Unknown Lighting” suggest that more lighting infrastructures are still needed on I-10 to ensure good visibility for truck drivers. On the other hand, freight truck should choose to operate during day time if possible which will both avoid the poor lighting conditions and fatigue driving behaviors.

For vehicle collision manners, compared with the “Single Vehicle” collision manner (the reference subcategory level), the “Rear End” and “Sideswipe Same Direction” are considered alleviating. This is consistent with the findings of researches (Chen and Chen, 2011; Kim et al., 2013) saying that the single vehicle accidents usually result in more serious injuries and such kind of collision resulted injuries account for 57.8% of all crash fatalities in 2005 and 46% of all motor vehicle fatalities in 2008 in the United States. The specific factors and prevention measures associated with the single vehicle crashes are beyond the research scope of this paper and interested readers should be referred to (Behnood and Mannering, 2015; Liu and Subramanian (2009); Wu et al., 2014).

Other identified critical factors include the violation of using electronic communication devices while driving, roadway design (road alignment and traffic way type), traffic control signal, drivers and passengers’ age, and the incident month. The study shows that “Using Electronic

Communication Device" is one contributing factor for severe injuries, thus the regulations for communication devices usage should be obeyed strictly for drivers and meanwhile more advanced and convenient devices should be equipped in trucks to mitigate the effects of this factor. For highway designers, the straight road alignment and the two way divided positive median barriers are encouraged to improve safety. For traffic control signal design, the flashing traffic signal (the reference level) is shown to improve the truck safety compared with the normal traffic signal lights on I-10 highway. For truck passengers older than 70 years old, the probability of being severely injured is greatly increased and in scenarios other than truck crashes researchers also identified this factor (Carter et al., 2014; Kim et al., 2013). Last, "September" is identified as alleviating for severe truck crash injuries and therefore more freight deliveries are recommended to be scheduled in September on I-10 Arizona segment from the safety perspective.

In sum, though many factors are identified as critical factors, the most counted critical factors (appeared more than 500 times in the Bagging model) are the damage area, the safety device, the severe crosswind weather and the high speed. Thus it is extremely beneficial to take strategies especially with respect to this 4 kinds of factors, meanwhile factors other than this 4 should also be taken seriously. Besides the non-spatial factors, the identified risky spatial factors should also cause more attention, such as the west rural area and urban area of Phoenix City, the south area of Tucson City, east of Benson City and so on. Traffic safety management is recommended to give those identified area priorities for safety improvement implementation.

3.5 Conclusion

This chapter provides a methodology framework of identifying and quantifying safety critical factors and applies it to the truck crash injury severity analysis. To handle the high dimensional

data case, modern statistical variable selection methods the Elastic Net and the Adaptive Elastic Net are utilized to improve the extensively used traditional logistic regression model. One advantage of this improvement is making the logistic regression model to incorporate thousands of factors which was previously not possible. But incorporating large amounts of factors can easily confuse model users who are looking for the most important factors leading to crashes. The modern variable selection methods can automatically shrink the coefficients of unimportant factors to zero while keeping the important ones. In such a way, the improved logistic regression model is built with sparse variables and meanwhile keeps the power for classification. Another contribution is to propose the Sub-Bagging technique to account for the imbalanced data structure which is common in traffic crash data especially for injury severity analysis. In the analysis of I-10 truck crash data, the relative risk interpretation is derived and used for critical factor identification and quantification. Last, based on the identified critical factors, safety improvement suggestions are come up and hopefully this will provide some beneficial guidance for highway traffic management.

4 Crash Accident Frequency Analysis

4.1 Introduction

The safety and mobility of a highway freight system are the 2 most concerned aspects for daily commuters. On one hand, the corridor's safety can be quantitatively assessed by the crash frequency, the number of crashes happened in the corridor over some specific period. A more detailed safety assessment provides the specific probabilities of crash incidents, addressing the questions such as "How likely will there be crash incidents?" and "What is the chance of the corridor having no crash accident?" On the other hand, the mobility is often described using the travel time and its uncertainty is critical for freight transportation decision making. The buffer time concept is proposed to account for such uncertainty and usually treated as representing travel time reliability. There is growing interest in examining the relationship between these 2 aspects. Especially for the safety analysis, few literatures have addressed the impacts of travel time reliability measures. This chapter seeks to provide a crash frequency assessment framework to understand such a relation, as well as the impacts of weather information. The analysis results could provide guidance for the corridor crash risk prediction and mitigation.

This chapter is organized as follows. The following section summarizes previous research work related to this study. Section 4.2 introduces the traffic data used in this research. Section 4.3 presents the modeling methods, including the spatial temporal model, mean function estimation, critical factor identification. Results interpretations and model validations are in Section 4.4. Section 4.5 summarizes this chapter.

4.2 Data Overview

The I-10 truck crash records of April and May in 2015 are used for this crash frequency investigation (ADOT, 2015). There are totally 490 truck crashes happened in the 2 months. In the following model development part, 42 days' crash numbers are used as training set and the rest 19 days' crash numbers are treated as test set for method validation purpose. The crash spatial distribution is summarized by Arizona counties through which that I-10 goes. Figure 4.1 highlights the differences of crash numbers among 5 counties. The crash temporal distribution is presented using the histogram in Figure 4.2, showing both the daily and Time of Day crash numbers.

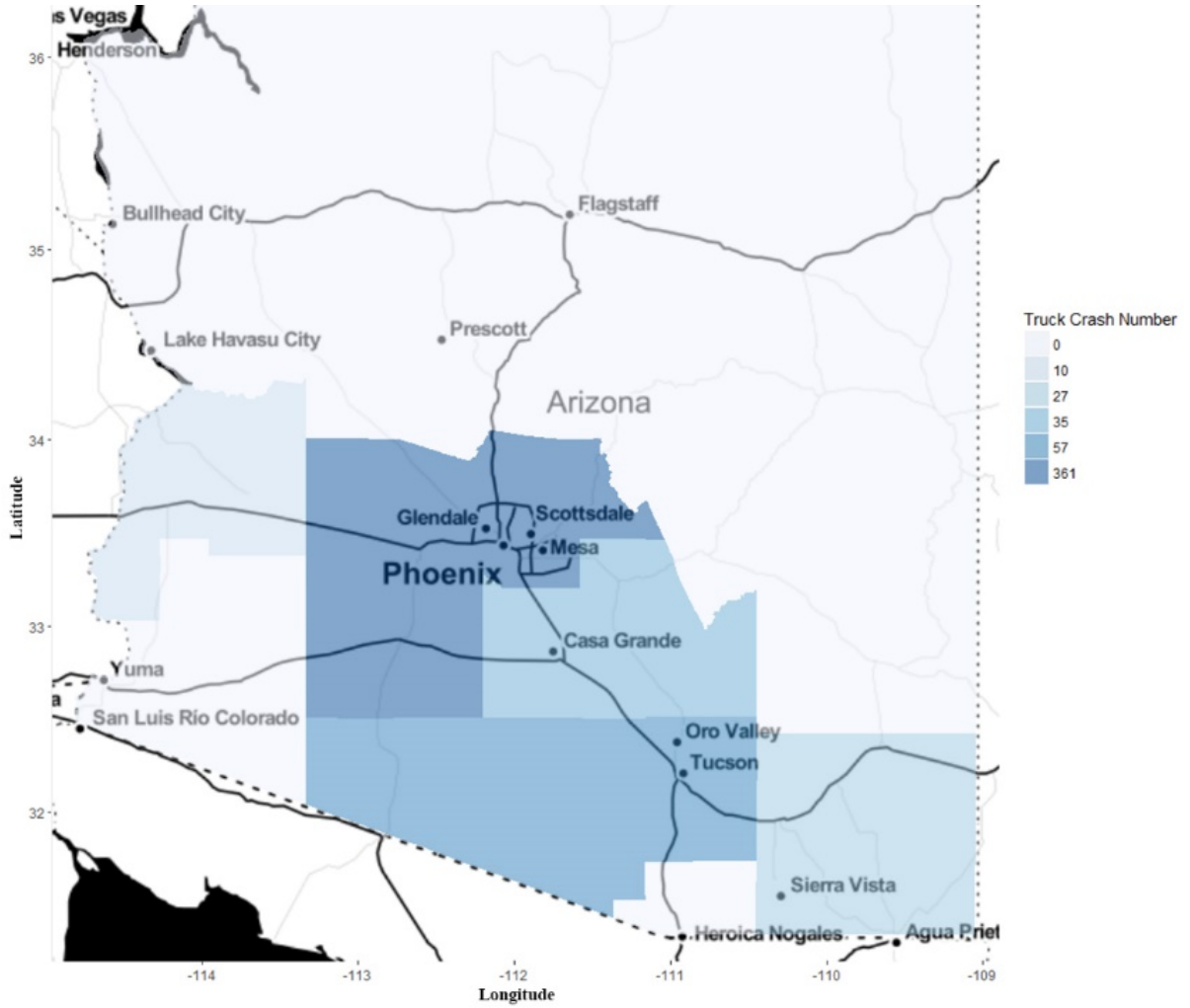


Figure 4. 1 Truck Crash Spatial Distribution

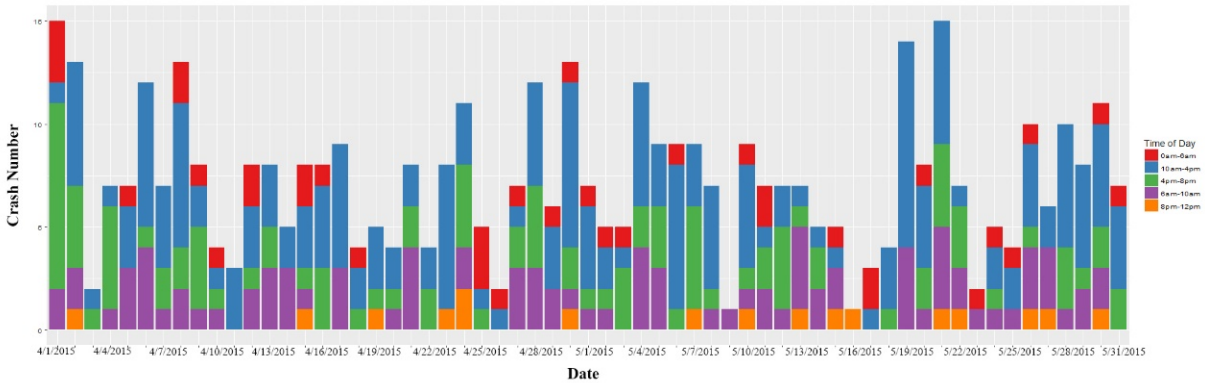


Figure 4. 2 Truck Crash Temporal Distribution

The travel time data are collected by the TMCs (Traffic Message Channel) which belong to the NPMRDS (National Performance Management Research Date Set) (FHWA, 2015). Each TMC covers a certain segment on highways and it records 5-minute average travel times of both westbound and eastbound trucks. The average length of each covered segment is 2.7 miles. The end GPS locations of TMC are also provided and can be used for visualization in map (Figure 4.3). There are 288 epoch records in total for a single TMC in one day and there are 147 TMCs along the I-10 corridor. Thus, for the whole corridor, there are 42336 records per day. Even though the data amount is huge, the data quality is not ideal and at least 20% of daily records are missing. Therefore, it is necessary to make remedies before the data can be utilized.

The weather data is provided by the underground database (WeatherUnderground, 2015) and it consists of the hourly temperature, wind speed and humidity information. 25 weather stations near the I-10 corridor are selected as potential weather factors that have impacts on truck crash frequency. The spatial distributions of TMCs and weather stations are presented in Figure 4.3.

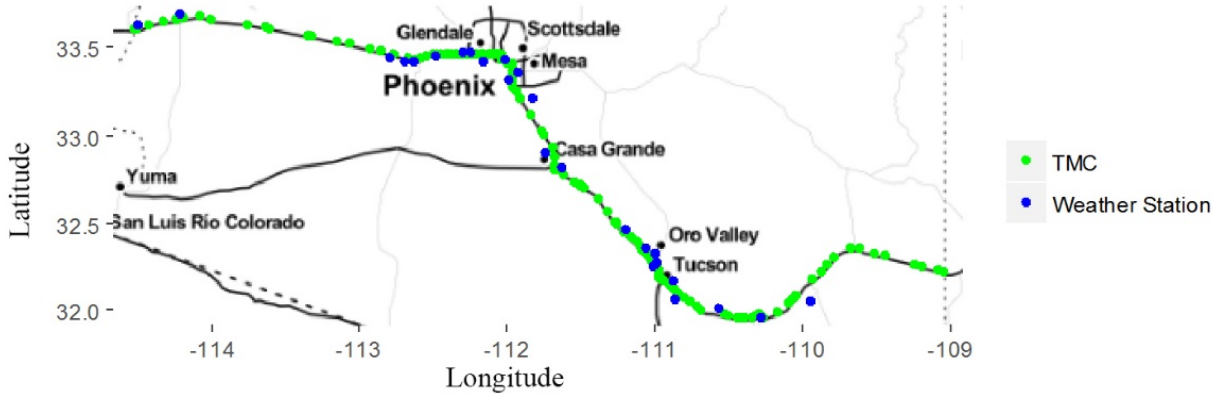


Figure 4. 3 TMC and Weather Station Spatial Distribution

4.3 Methodology

A crash frequency analysis framework focused on investigating impacts of travel time reliability and weather conditions is proposed in this paper. Different from the traditional Safety Performance Function (SPF) and other aggregated data (e.g., yearly or monthly) based methods (Kononov et al., 2008; Wang et al., 2009), the proposed method utilizes the high resolution travel time and weather data, aiming to capture important potential time varying factors leading to crash incidents. First, a spatial temporal model is introduced and applied for travel time estimation, considering observation noise and missing value conditions. Then, by examining the heterogeneity of mixed Poisson regressions, a common mean function is developed for the time-of-day crash numbers based on all potential factors. Next, critical segments and weather observation stations are identified through modern statistical variable selection methods. Instead of treating segment traffic conditions independently, the identification procedure takes correlations among factors into consideration. Marginal effects analysis is implemented for the interpretation purpose. Last, to account for crash risk quantification, a formal statistical test is performed for distribution selection. Overall, this data analytics framework provides the possibility to conduct crash risk assessment and prediction under the condition that the data quality is not ideal and the number of potential factors is huge. Before proceeding to the next section, a notation table (Table 1) is included for reference

Table 4. 1 Notation List

$Y(\boldsymbol{s};t)$	variable indicating travel time of location **s** at time t	μ_i	mean parameter in Poisson regression
$C(\mathbf{s}-\mathbf{z};t-r)$	covariance function between travel time of location **s** at time t and of location **z** at time r	ν_i	heterogeneity term in Poisson regression
$\mathbf{x}_t$	travel time vector at time t	σ, ν	parameters defines heterogeneity term's PDF
$\mathbf{y}_t$	observed travel time vector at time t	$l(\bullet)$	link function in Generalized Linear Model
Φ	evolution matrix for travel time	η_i	transformed mean parameter by the link function
$\mathbf{w}_t$	vector indicating randomness of state process	$\boldsymbol{\beta}$	coefficients of explanatory variables
Q	covariance matrix of state randomness	$P_\alpha(\boldsymbol{\beta})$	penalty term for coefficient estiamtion
$\mathbf{v}_t$	vector indidcating travel time observation noise	λ	weight for penalty term
R	covariance matrix of observation noise	α	weight for absolute value of coefficient in penalty term
A_t	travel time measurement matrix	x_{ij}^R	recovered travel time at location j at time t in spatial temporal model validation
$\mathbf{y}_t^1$	observed value at time t in incomplete observed travel time vector	x_{ij}^B	benchmark method recovered travel time in spatial temproal model validation
$\mathbf{y}_t^0$	missing value at time t in incomplete observed travel time vector	y_{MAE}	Mean Absolute Error (MAE) of certain factor level combinations in spatail temporal model validation
A_t^1	travel time measurement matrix related to observed values	μ_{MAE}	grand mean of MAE
A_t^0	travel time measurement matrix related to missing values	τ_i	Day of Week effects on MAE in spatial temporal model validation
$\mathbf{v}_t^1$	travel time observation noise related to observed values	γ_j	Time of Day effects on MAE in spatial temporal model validation
$\mathbf{v}_t^0$	travel time observation noise related to missing values	β_k	Location effects on MAE in spatial temporal model validation
R_{11t}	covariance matrix of noise related to observed values	η_m	Direction effects on MAE in spatial temporal model validation
R_{22t}	covariance matrix of noise related to missing values	υ_n	imputation method effects on MAE in spatial temporal model validation
R_{12t}, R_{21t}	covariance matrix of noise between observed and missing values	N_T	temporal range of MAE calculation in spatial temporal model validation
$\boldsymbol{x}_t^n$	expected travel time given n available observations	N_L	spatial range of MAE calculation in spatial temporal model validation
$P_{t,t-1}^n, P_t^n$	covariance matrix of travel time vectors	ε	white noise in spatial temporal model validation
$Q\left(\Theta \mid \Theta^{(j-1)}\right)$	expected conditional joint likelihood at iteration j	$\overline{p}_j$	crash frequency computed from the crash sample
y_i^C	daily time-of-day crash number	$\hat{p}_j^F$	crash frequency computed from fitted regressions
$\boldsymbol{x}_i^C$	explanatory variable vector related to crash		

4.3.1 Spatial Temporal Model for Travel Time

(1) State Space Model

To model the spatial temporal relation of a continuous process, one common approach is the descriptive approach characterizing the first and the second moment behavior of the process (Cressie and Wikle, 2015). Assuming at time t the travel time at location $\boldsymbol{s}$ is denoted by $Y(\mathbf{s};t)$,

then the disadvantage of this approach is that it assumes the stationarity of the process which demands: the constant expectation of $Y(\boldsymbol{s};t)$ across the domain of interest and the spatial temporal covariance functions is only depending on the separation in spatial and temporal dimensions:

$$\operatorname{cov}\left(Y(\boldsymbol{s};t),Y(\boldsymbol{z};r)\right)=C(\boldsymbol{s}-\boldsymbol{z};t-r);\boldsymbol{s},\boldsymbol{z}\in\mathbb{R}^{d},t,r\in\mathbb{R} \tag{4.1}$$

In some applications, stronger conditions are imposed to simplify the calculations such as the isotropy property and the separable property. By the Isotropy, the covariance function is simplified to $C(\|\boldsymbol{s}-\boldsymbol{z}\|;|t-r|)$ and by the Separable property, it is simplified to $C^{(s)}(\boldsymbol{s},\boldsymbol{x})C^{(t)}(t,r)$ in which $C^{(s)}$ and $C^{(t)}$ are respectively spatial and temporal covariance functions. In the modelling of travel time spatial temporal relations, it is not appropriate to consider the travel time has the stationary property, since it is well known that the traffic conditions vary significantly throughout the day, particularly when the rush hours and non-rush hours are taken into consideration. On the other hand, the interactions between the spatial influence and temporal influence should be included in modelling. Therefore, another way for modeling the spatial temporal relationship is come up with, named the dynamic approach. Specifically, we propose using the state space modelling approach. The Dynamic Linear Model (DLM) consists of two parts, the state equation and the observation equation.

$$\boldsymbol{x}_t = \Phi\boldsymbol{x}_{t\text{-}1} + \boldsymbol{w}_t \tag{4.2}$$

$$\boldsymbol{y}_t = A_t\boldsymbol{x}_t + \boldsymbol{v}_t \tag{4.3}$$

In which $\boldsymbol{x}_t$ is the $p \times 1$ state vector at time $t = 1,2,\cdots,n$ and it contains p real travel time at TMC detectors locations. It assumes that the process start with a normal vector $\boldsymbol{x}_0 \sim N_p(\mu_0, \Sigma_0)$. The $p \times p$ transition matrix Φ is indicating the spatial and temporal interaction relations in travel time evolutions. The randomness of the state process is represented by the $p \times 1$ vector $\boldsymbol{w}_t \sim i.i.d. N_p(\boldsymbol{0}, Q)$. In the observation equation, $\boldsymbol{y}_t$ is the $p \times 1$ direct observed travel time data vector at time t recorded by all the p TMC detectors and A_t is the $p \times p$ measurement matrix. However, it is assumed the data is observed with additional noise $\boldsymbol{v}_t \sim i.i.d. N_p(\boldsymbol{0}, R)$. For travel time observations, the measurement matrix A_t is an identity matrix since each observation is assumed to have the form:

$$y_{ti} = x_{ti} + v_{ti}, i \in \{1, 2, \cdots, p\} \tag{4.4}$$

Given the formulation of the state space model, the Kalman filter and smoother is considered more efficient algorithms for travel time state estimation.

(2) Parameter Estimation

To efficiently estimate unknown parameters $\{\Phi, Q, R\}$, the Expectation Maximization (EM) method is more robust and preferred than the commonly implemented quasi-Newton algorithms

(i.e. the Broyden-Fletcher-Goldfarb-Shanno (BFGS)), under the condition that there is large amount of missing observations (Holmes et al., 2013). The observation equation and its covariance matrix of measurement noises are rewritten as:

$$\begin{pmatrix} \boldsymbol{y}_t^1 \\ \boldsymbol{y}_t^0 \end{pmatrix} = \begin{pmatrix} A_t^1 \\ A_t^0 \end{pmatrix} \boldsymbol{x}_t + \begin{pmatrix} \boldsymbol{v}_t^1 \\ \boldsymbol{v}_t^0 \end{pmatrix} \tag{4.5}$$

$$\operatorname{cov}\begin{pmatrix} \boldsymbol{v}_t^1 \\ \boldsymbol{v}_t^0 \end{pmatrix} = \begin{bmatrix} R_{11t} & R_{12t} \\ R_{21t} & R_{22t} \end{bmatrix} \tag{4.6}$$

Where the superscript 1 and 0 are indicating if the component is related to the observed value or the missing value. R_{12t} and R_{21t} are both assumed to be zero if the observed and the unobserved have uncorrelated errors. Under the condition of the missing observations, the state and covariance estimators are denoted as:

$$\boldsymbol{x}_t^n = E\left(\boldsymbol{x}_t \mid \boldsymbol{y}_1^1, \boldsymbol{y}_2^1, \cdots \boldsymbol{y}_n^1\right) \tag{4.7}$$

$$P_{t,t-1}^n = E\left[\left(\boldsymbol{x}_t - \boldsymbol{x}_t^n\right)\left(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^n\right)^T\right] \tag{4.8}$$

$$P_t^n = E\left[\left(\boldsymbol{x}_t - \boldsymbol{x}_t^n\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_t^n\right)^T\right] \tag{4.9}$$

The state and covariance estimates (4.7, 4.8, 4.9) are derived from Kalman filtering and smoothing recursions (Shumway and Stoffer, 2010). The details of Kalman Filter and Smoother

are in Appendix B. The EM is an iterative algorithm and it has two major steps, the E (Expectation) step and the M (Maximization) step. In the E step, the expectation of joint likelihood is calculated using current parameters. In the M step, the parameters are updated through the multivariate normal maximum likelihood estimators. The procedure is:

(1) Initialize the unknown parameters $\Theta^{(0)} = \left\{ \boldsymbol{\mu}_0^{(0)}, \Sigma_0^{(0)}, \boldsymbol{\Phi}^{(0)}, Q^{(0)}, R^{(0)} \right\}$

(2) The E step: given $\Theta^{(j-1)}$ obtained iteration $j-1$, calculate:

$$Q\left(\Theta \mid \Theta^{(j-1)}\right) = E\left\{-2\ln L_{X,Y}\left(\Theta\right) \mid \boldsymbol{y}_1^1, \boldsymbol{y}_2^1, \cdots \boldsymbol{y}_n^1, \Theta^{(j-1)}\right\} \tag{4.10}$$

In which, given the joint density:

$$f_\Theta\left(\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots \boldsymbol{x}_n, \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots \boldsymbol{y}_n\right) = f_{\boldsymbol{\mu}_0, \Sigma_0}\left(\boldsymbol{x}_0\right) \prod_{t=1}^{n} f_{\boldsymbol{\Phi}, Q}\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) \prod_{t=1}^{n} f_R\left(\boldsymbol{y}_t \mid \boldsymbol{x}_t\right) \tag{4.11}$$

The joint likelihood is:

$$\begin{aligned} -2\ln L_{X,Y}\left(\Theta\right) &= \ln\left|\Sigma_0\right| + \left(\boldsymbol{x}_0 - \boldsymbol{\mu}_0\right)^T \Sigma_0^{-1}\left(\boldsymbol{x}_0 - \boldsymbol{\mu}_0\right) + n\ln\left|Q\right| + \\ &\sum_{i=1}^{n}\left(\boldsymbol{x}_t - \boldsymbol{\Phi}\boldsymbol{x}_{t-1}\right)^T Q^{-1}\left(\boldsymbol{x}_t - \boldsymbol{\Phi}\boldsymbol{x}_{t-1}\right) + n\ln\left|R\right| + \sum_{i=1}^{n}\left(\boldsymbol{y}_t - A_t\boldsymbol{x}_t\right)^T R^{-1}\left(\boldsymbol{y}_t - A_t\boldsymbol{x}_t\right) \end{aligned} \tag{4.12}$$

(3) The M step: minimize $Q\left(\Theta \mid \Theta^{(j-1)}\right)$ with respect to Θ and set $\Theta^{(j)}$ equal to the minimizer:

$$\boldsymbol{\Phi}^{(j)} = S_{10}S_{00}^{-1} \tag{4.13}$$

$$Q^{(j)} = n^{-1}\left(S_{11} - S_{10}S_{00}^{-1}S_{10}^{T}\right) \tag{4.14}$$

$$R^{(j)} = n^{-1}\sum_{t=1}^{n} D_t\left[\left(\boldsymbol{y}_t - \boldsymbol{A}_t\boldsymbol{x}_t^n\right)\left(\boldsymbol{y}_t - \boldsymbol{A}_t\boldsymbol{x}_t^n\right)^T + \boldsymbol{A}_t P_t^n \boldsymbol{A}_t^T + \begin{bmatrix} 0 & 0 \\ 0 & R_{22t}^{(j-1)} \end{bmatrix}\right] D_t^T \tag{4.15}$$

In which, D_t is the permutation matrix that reorders the variables at time *t* and S_{00}, S_{10}, S_{11} are defined as:

$$S_{00} = \sum_{t=1}^{n}\left(\boldsymbol{x}_{t-1}^n\left(\boldsymbol{x}_{t-1}^n\right)^T + P_{t-1}^n\right) \tag{4.16}$$

$$S_{10} = \sum_{t=1}^{n}\left(\boldsymbol{x}_{t}^n\left(\boldsymbol{x}_{t-1}^n\right)^T + P_{t,t-1}^n\right) \tag{4.17}$$

$$S_{11} = \sum_{t=1}^{n}\left(\boldsymbol{x}_{t}^n\left(\boldsymbol{x}_{t}^n\right)^T + P_{t}^n\right) \tag{4.18}$$

(4) Stop if the expectation of negative joint likelihood decreases less than 0.01; otherwise increase j by 1 and return to the E step.

4.3.2 Crash Risk Model

(1) Heterogeneity in Poisson Regression

The crash data is essentially count data. The Poisson regression is the most implemented method by conditioning the mean on predictors, which allowing different individuals to have different Poisson means. However, there are still unexplained heterogeneity leading to departures from the Poisson process assumptions. Cameron and Trivedi (2013) introduced some common departures including the overdispersion which is defined as the extra variation occurred in modeling count data, not explained by the Poisson distribution alone. Usually, the unobserved heterogeneity is introduced as a multiple of Poisson mean. Let $f(y_i^C \mid \mu_i, v_i)$ denote the Poisson Probability Mass Function (PMF) with $\mu_i v_i$ the mean parameter and y_i^C is the daily time-of-day observed crash count, where the subscript i denotes the individual differences among the sample and the superscript C is used to differentiate the observed travel time in Section 4.3.1. The PMF becomes:

$$f(y_i^C \mid \mu_i, v_i) = \frac{e^{-\mu_i v_i} (\mu_i v_i)^{y_i}}{y_i!} \tag{4.19}$$

If a further Probability Density Function (PDF) for v_i is denoted by $g(v_i)$, then the marginal density of $y_i^C \mid \mu_i$ can be derived as:

$$h(y_i^C \mid \mu_i) = \int f(y_i^C \mid \mu_i, v_i) g(v_i) dv_i \tag{4.20}$$

Meanwhile, the heterogeneity term v_i is assumed to be independently identically distributed with properties $E[v_i \mid \mu_i] = E[v_i]$ and $E[v_i] = 1$. Under these conditions, by some simple algebra it can be shown $E\left[y_i^C \mid \mu_i\right] = \mu_i$. The class of mixture models are derived using different distribution assumptions about the heterogeneity term, but the marginal distributions of counts are all based on the same mean parameter μ_i as the original Poisson regression has. Here, we follow the work of Rigby et al. (2008) and the mixed Poisson distributions are summarized by their heterogeneity term v_i's PMF. The Poisson distribution with only one mean parameter strictly requires that the mean equals the variance, whereas the rest distributions allow more skewness (Table 4.2). Further, the three-parameter defined distributions (the Sichel distribution and the Delaporte distribution) have more flexible skewness-kurtosis combinations than the two-parameter defined distributions (the Negative Binomial distribution, the Poisson Inverse Gaussian distribution and the Zero Inflated Poisson distribution).

Table 4. 2 Mixed Poisson Distributions

Heterogeneity Distribution	Marginal Distribution	Mean	Variance
NA	Poisson (μ)	μ	μ
Gamma ($1,\sigma^{1/2}$)	Negative Binomial Type I (μ,σ)	μ	$\mu+\sigma\mu^2$
Gamma ($1,\sigma^{1/2}/\mu$)	Negative Binomial Type II (μ,σ)	μ	$\mu+\sigma\mu$
Inverse Gaussian ($1,\sigma^{1/2}$)	Poisson Inverse Gaussian (μ,σ)	μ	$\mu+\sigma\mu^2$
$(1-\sigma)^{-1}$Binomial (1,1-σ)	Zero Inflated Poisson (μ,σ)	μ	$\mu+(1-\sigma)^{-1}\sigma\mu^2$
Generalized Inverse Gaussian ($1,\sigma^{1/2},\nu$)	Sichel (μ,σ,ν)	μ	$\mu+h(\sigma,\nu)\mu^2$
Shiftted Gamma ($1,\sigma^{1/2},\nu$)	Delaporte (μ,σ,ν)	μ	$\mu+(1-\nu)^2\mu^2$

(2) Critical Variable Identification

Since the mixed Poisson models are sharing the same mean parameter, it is convenient to build different heterogeneity forms given the fixed mean parameter and the derivation of the mean function is irrelevant to the distribution assumption. This class of models belongs to the Generalized Linear Model (GLM) family in which the linear predictor η_i for the count response y_i^C is based on explanatory variables $\boldsymbol{x}_i^C = \left(x_{i1}^C, x_{i2}^C, \cdots, x_{iq}^C\right)^T$, where q denotes the number of explanatory variables (Nelder and Baker, 1972a). The relation is:

$$l\left(\mu_i\right) = \eta_i = \left(\boldsymbol{x}_i^C\right)^T \boldsymbol{\beta} \tag{4.21}$$

In which $\boldsymbol{\beta}$ is the $q \times 1$ coefficient vector. The function $l(\bullet)$ is called the link function relating μ_i to η_i. In modelling the crash data, the log link function is implemented to ensure the predicted mean μ_i to be non-negative. The specification for the mean function is:

$$\mu_i = \exp\left(\left(\boldsymbol{x}_i^C\right)^T \boldsymbol{\beta}\right) \tag{4.22}$$

The explanatory variables are representing the locations' one-mile buffer time, the temperature, the wind speed and the humidity measures. Traditionally, the buffer time is defined by using the difference between the 95th percentile travel time and the average travel time. Some researchers suggest using the median travel time instead (replacing the average with the median) and claim

the resulted median based buffer time is superior to the average-based (Pu, 2011). Here we are using the latter definition of buffer time, since it is more robust to some abnormal travel time records. Moreover, since our purpose is to investigate the impacts of buffer time and to avoid the "cause and effect" ambiguity, the one hour travel time right after crash incidents are excluded in buffer time calculations.

Considering one direction of I-10 corridor, there are 147 detectors. Meanwhile, there are 25 weather observation stations along I-10 and each of them measures the temperature, the wind speed and the humidity parameters. Thus, the explanatory vector $\boldsymbol{x}_i^C$ is a 223×1 vector including the constant intercept term. The directly estimation of $\boldsymbol{\beta}$ using Least Square methods and the variance based variable significance tests are not possible. Because the number of coefficients is greater than the sample size (42 days) and the $q\times q$ matrix $\left(X^C\right)^T X^C$ is not invertible. Many modern variable selection methods were come up including the LASSO (Least Absolute Selection and Shrinkage Operator), the SCAD (Smoothly Clipped Absolute Deviation), the Nonnegative Garrote and more. To estimate the coefficients and select the significant buffer time, the Elastic Net (EN) variable selection method is implemented. The advantages of this EN variable selection method are its additivity to high dimensionality and its ability to select highly correlated grouped covariates (Zou and Hastie, 2005). In this research, coefficients of explanatory variables are estimated by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i^C - \mu_i\right)^2 + \lambda P_\alpha\left(\boldsymbol{\beta}\right) \tag{4.23}$$

In which the penalty term $P_\alpha(\boldsymbol{\beta})$ is:

$$P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\|\boldsymbol{\beta}\|^2 + \alpha\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{q}\left[(1-\alpha)\beta_j{}^2 + \alpha|\beta_j|\right], \alpha \in (0,1) \quad (4.24)$$

This optimization problem can be solved by the Cyclical Coordinate Descent Algorithm (Friedman et al., 2010). Tuning parameters λ and α are determined by a 2 dimensional 10-fold Cross Validation. The purpose of the imposed penalty terms is to improve the prediction accuracy and reduce the complexity of the conditional mean function by soft thresholding on β_j s among which small coefficients are shrunk to zeros. Thus, a sparse $\hat{\boldsymbol{\beta}}$ is obtained and the remained non-zero coefficients are representing travel time and weather condition effects of our interests.

4.4 Results and Model Validation

The analysis results are presented in terms of travel time imputation, critical variable selection and crash risk distribution selection. To validate these results, designed experiments and formal tests are conducted. In Section 4.4.1, examples of travel time imputation are presented. Then, to find out how good the recovered records are, which factors are affecting the imputation accuracy and how these factors are different from each other, an experiment of factorial design is implemented. In Section 4.4.2, critical segments and weather stations are identified by the EN variable selection procedure and their effects on expected crashes are quantified. Further, crash number prediction errors both in training sets and test sets are reported, indicating that the identified segments and

stations have satisfied prediction capability. In Section 4.4.3, to correctly and quantitatively assess the crash risk, a group of candidate mixed Poisson regression models are compared using the Chi-square Goodness of Fit test.

4.4.1 Spatial Temporal Model

The missing values of daily 5-minute epoch travel time $\boldsymbol{x}_t$ are imputed by Kalman smoother estimators $\boldsymbol{x}_t^n$. An intuitive illustration is presented using one day's travel time data (Figure 4.4). The 3 dimension pictures are showing eastbound raw travel time records and Kalman smoother recovered records on Apr. 15 in 2015. The X axis represents the epoch time (288 5-minute epochs for one day) and the Y axis represents the TMC (147 in total). These 2 pictures are essentially matrix visualization and each matrix element represents the 5-minute average travel time in seconds for one mile.

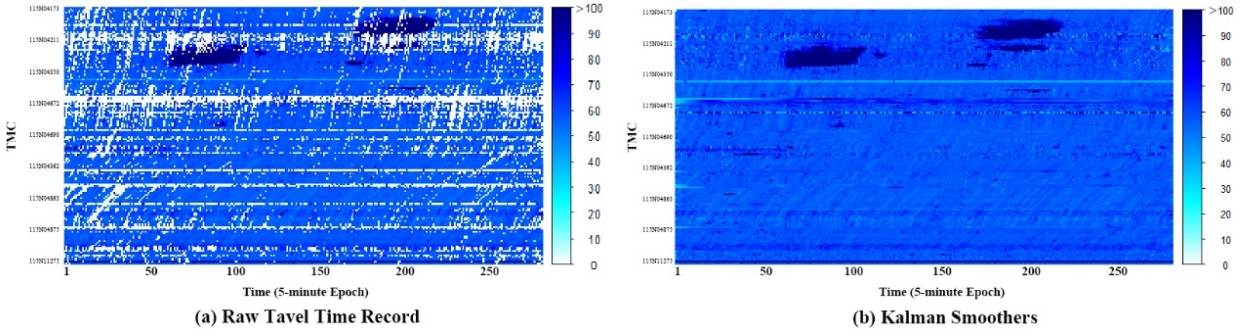


Figure 4. 4 Eastbound Travel Time Imputation on Apr. 15 of 2015

To assess the quality of the recovered values under the incomplete data condition, an experiment of multiple factorial design is conducted (Montgomery, 2005). 20% of daily observed raw travel

time records are randomly selected and removed. The removed 20% would be imputed by Kalman smoothers and the average of daily records respectively, using the rest 80% of available observations. Meanwhile the benchmark Kalman smoother estimators are derived without removing any data from original datasets. The comparison is made between the values recovered by 80% raw data and the ones recovered by 100% raw data. The Mean Absolute Error (MAE) is used as the criterion for accuracy assessment.

$$MAE = \frac{1}{N_T N_L} \sum_{i=1}^{N_T} \sum_{j=1}^{N_L} \left| x_{ij}^R - x_{ij}^B \right| \tag{4.25}$$

Where x_{ij}^R is the recovered jth location's travel time at time i and x_{ij}^B is the benchmark method recovered travel time. N_T and N_L are indicating the temporal and spatial ranges of this defined MAE measure. The statistical model of this experiment is:

$$y_{MAE} = \mu_{MAE} + \tau_i + \gamma_j + \beta_k + \eta_m + \upsilon_n + \varepsilon \tag{4.26}$$

In which y_{MAE} is the MAE value of specific factor level combinations and μ_{MAE} is the grand mean MAE of the whole recovered travel time. $\{\tau_i, \gamma_j, \beta_k, \eta_m, \upsilon_n\}$ are the effects of Day of Week factor, Time of Day factor (different daily rush hours and non-rush hours), the location factor (Arizona counties), the direction factor and the imputation method factor (the Kalman smoother and the

average). The subscripts are indicating their own specific subcategory level ranges which can be found in Table 4.4. ε is the white noise accounting for extra randomness in MAE. Since all the 5 factors have fixed levels, a simple Analysis of Variance (ANOVA) is implemented (with the significance level $\alpha = 0.05$) showing all the factors are significant with respect to the imputation accuracy (Table 4.3). The Tukey's pairwise comparison (with the significance level $\alpha = 0.05$) shows the differences of MAE among effect levels belonging to the same factor (Table 4.4). Levels with same group letters are considered not significantly different. According to the Tukey's pairwise comparison table, it is found that the recovered 5 minute epoch travel time using incomplete data is most accurate on Saturday and Sunday with an MAE value around 4 seconds, in terms of the Day of Week factor. For the Time of Day factor, the time periods 8pm to 12pm and 0am to 6am are found to have smaller MAE values than other periods. Among the 5 counties the highway goes through, La Paz County and Cochise County have more accurate recovered travel time relatively. It is also noticed that the imputed eastbound travel time has smaller MAE than the imputed westbound travel time. As for the method factor, the Kalman smoother is shown to be better than using the daily average for imputation.

Table 4. 3 ANOVA

Factor	Sum of Squares	Degrees of Freedom	Mean Square	F Value	P Value
Day of Week	20623	6	3437	131.99	2×10^{-16}
Time of Day	13673	4	3418	131.26	2×10^{-16}
Location	39282	4	9820	377.10	2×10^{-16}
Direction	2131	1	2131	81.83	2×10^{-16}
Method	2606	2	2606	100.08	2×10^{-16}
Residuals	158413	6083	26		

Table 4. 4 Tukey's Pairwise Comparison Result

Factor	Level	MAE	Group	Factor	Level	MAE	Group
Day of Week	Tuesday	8.70	a_1	Location	Pinal County	10.79	a_3
	Thursday	8.41	a_1b_1		Maricopa County	8.75	b_3
	Wednesday	7.79	b_1c_1		Pima County	5.09	c_3
	Monday	7.62	c_1		Cochise County	4.86	c_3d_3
	Friday	7.17	c_1		La Paz County	4.39	d_3
	Saturday	4.03	d_1				
	Sunday	4.01	d_1	Direction	West	7.37	a_4
					East	6.18	b_4
Time of Day	4pm-8pm	9.48	a_2	Method			
	6am-10am	7.29	b_2		Average	8.43	a_5
	10am-4pm	6.01	c_2		Kalman Smoother	6.12	b_5
	0am-6am	5.67	c_2d_2				
	8pm-12pm	5.42	d_2				

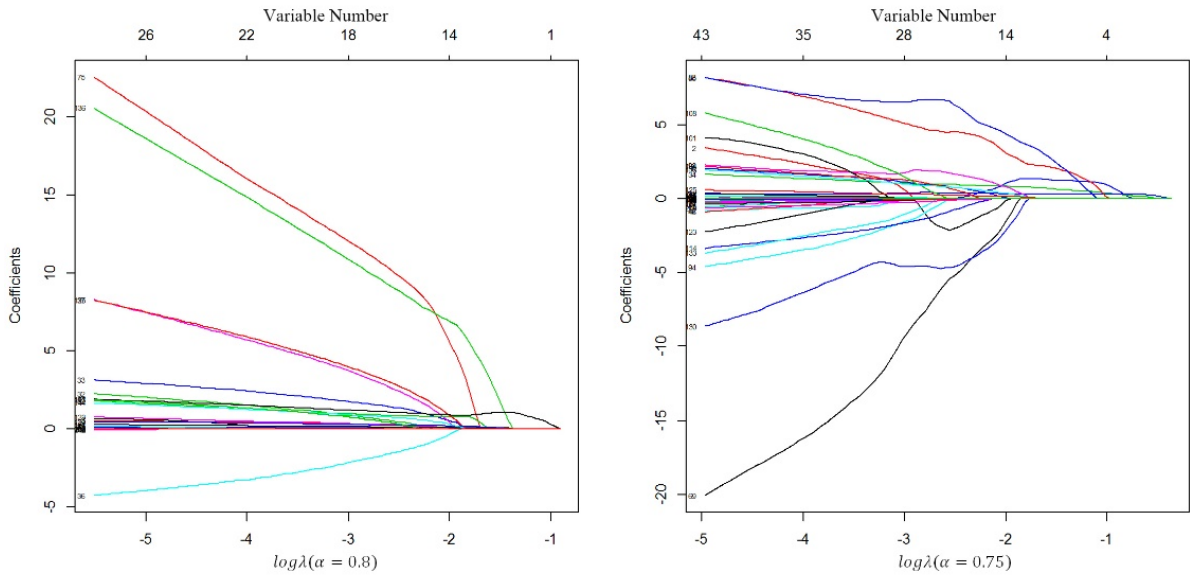
4.4.2 Critical Variables

(a) Model Validation

Selected critical segments and weather stations are different for each period time of the day. The variable selection process follows a continuous process. Take the case of 0am-6am for example, shrinkage paths of variable coefficients are presented in Figure 4.5. By a 10-fold Cross Validation, turning parameters are identified as $\lambda = 0.15, \alpha = 0.8$ for the westbound mean function and $\lambda = 0.09, \alpha = 0.75$ for the eastbound mean function. The MAE is used to quantify the differences between the predicted truck crash numbers and the real numbers of crashes.

$$MAE = \frac{1}{n_I}\sum_{i=1}^{n_I}\left|\mu_i - y_i^C\right| \tag{4.27}$$

Where the subscript $I \in \{0,1\}$ is indicating whether the sample is from the training or test dataset. The prediction results are reported in Table 4.5 including both the training errors and the test errors. Notice that using the mean function only gives a rough prediction and better prediction should be combined with specific distributions. For example, the mean function usually gives a number with decimal numbers which are meaningless in describing crash numbers and people usually prefer integer numbers when talking about 'How many crashes are going to happen given current traffic conditions and weather?' Thus, this mean function prediction only serve to validate the effeteness of the variable selection method and the MAE measure shows the raw magnitude of errors. More meaningful predictions and performance measures are proposed in Section 4.5 after certain distribution regressions are fitted.



(a) Westbound (b) Eastbound

Figure 4. 5 0am-6am Mean Function Variable Shrinkage Paths

Table 4. 5 Mean Function Prediction Error

	0am-6am	6am-10am	10am-4pm	4pm-8pm	8pm-12pm
Westbound Training Error	0.29	0.47	0.59	0.3	0.23
Westbound Test Error	0.49	0.67	1.08	0.66	0.33
Eastbound Training Error	0.41	0.53	0.59	0.68	0.14
Eastbound Test Error	0.82	1.16	0.91	0.98	0.15

(b) Interpretation

The daily period crash numbers are estimated through formula (28). Even though the relation between the explanatory variables (buffer time and weather) and crash number is established, it is not straightforward to understand the meaning of GLMs coefficients because of the link function. Therefore, the Average Marginal Effect (AME) is employed for the interpretation purpose and the definition is:

$$AME_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mu_i}{\partial x_{ij}^C} = \frac{1}{n}\sum_{i=1}^{n}\beta_j \exp\left(\left(\boldsymbol{x}_{ij}^C\right)^T \boldsymbol{\beta}\right) \quad (4.28)$$

The AME_j is interpreted as the effect of a one unit increase in the explanatory variable $x_{\cdot j}^C$ on the expected truck crash number in the investigated period of the day. For buffer time variables, the unit is the minute rather than the second, since the AME of a one second change in buffer time is too subtle. The identified critical segments, weather stations and their AME are listed in Table 4.6 and Table 4.7. The positive AME values indicate the identified segment's increased buffer time are contributing to the expected crash mean number. For those segments with negative AME values, their increased buffer time is likely to lower the crash number. This could be explained by that fewer severe crashes are expected to occur when the traffic condition degrades (Zhou and Sisiopiku, 1997). On the other hand, most weather-related variables have positive AME values. For example, all the wind speed variables have positive AME values, indicating the higher wind speed is likely to increase the crash number.

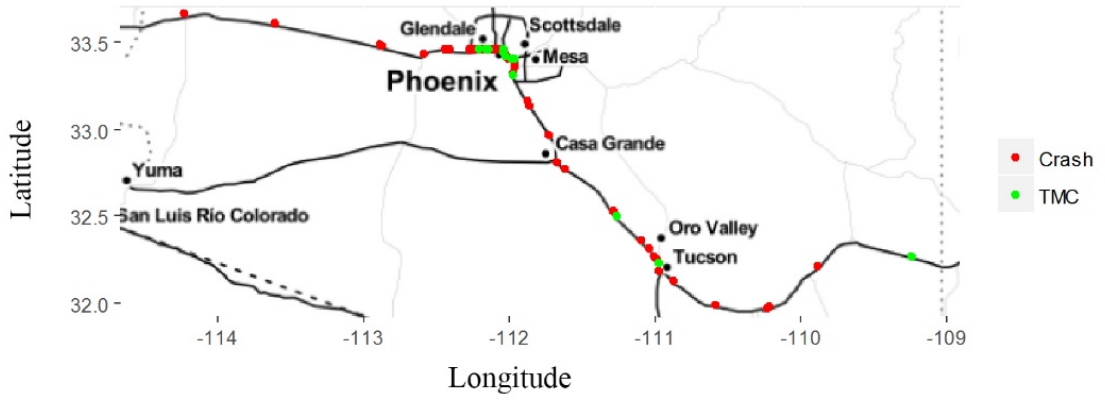
To illustrate the spatial distribution of those identified segments and weather stations, Figure 4.6 is presenting a time of day case (10am-4pm). It is easy to see that most of the truck crashes happened around the identified critical TMC detectors and weather stations. In fact, the median distance between a crash and the nearest end of identified TMC segment or Weather Station is 2.8 miles almost as far as a TMC detector's coverage.

Table 4. 6 Identified Critical Segment's TMC and AME

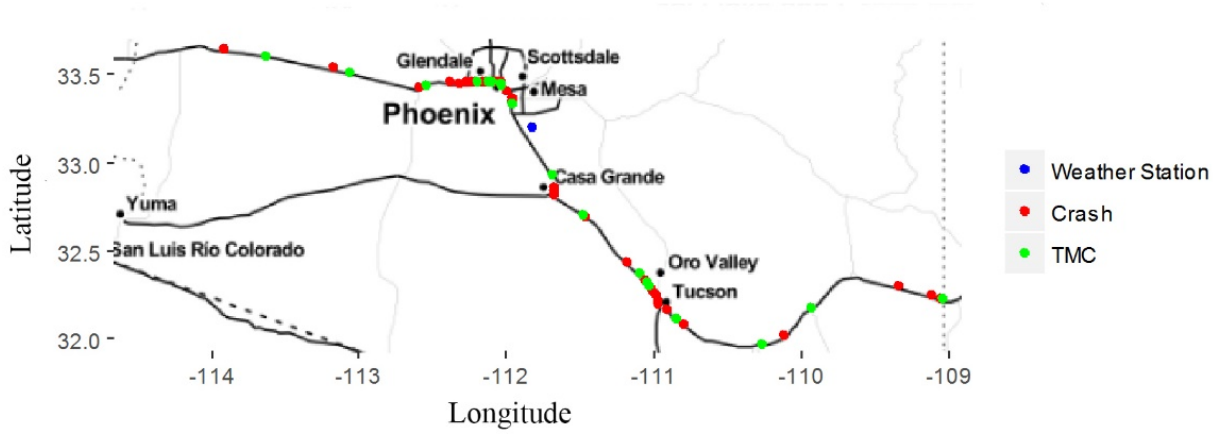
Direction	0am-6am		6am-10am		10am-4pm		4pm-8pm		8pm-12pm	
	TMC	AME	TMC	AME	TMC	AME	TMC	AME	TMC	AME
Westbound	115N04888	2.700	115N04675	1.072	115N04185	0.180	115N04350	0.930	115N04859	0.091
	115N04672	1.285	115N04201	0.079	115N04188	0.107	115N04753	0.147	115N04877	0.043
	115N04877	0.274	115N04181	0.045	115N04184	0.065	115N04205	0.093	115N04200	0.042
	115N04205	0.123	115N04200	0.041	115N04859	0.044	115N04189	0.063	115N04197	0.026
	115N04668	0.082	115N04342	0.032	115N04668	0.041	115N04206	0.055	115N04862	0.015
	115N04679	0.076	115N04179	0.022	115N04192	0.036	115N04191	0.030	115N04194	0.001
	115N04875	0.050	115N04182	0.018	115N04354	0.020	115N04202	0.026		
	115N04891	0.043	115N04671	0.016	115N04201	0.017	115N04194	0.021		
	115N04175	0.038	115N04180	0.012	115N04202	0.016	115N04875	0.019		
	115N04208	-0.125	115N04856	0.012	115N04193	0.011	115N04199	0.002		
			115N04343	0.012	115N04205	0.007	115N04749	-1.132		
			115N04184	0.009	115N04178	0.003				
					115N04194	0.000				
Eastbound	115N04181	0.131	115N04205	0.048	115N04361	3.132	115N07208	0.164	115N04690	0.210
	115N04207	0.148	115N04204	0.022	115N04689	0.391	115N07209	0.049	115N04866	0.048
	115N04206	0.217	115N04340	0.015	115N04360	0.320	115N04679	0.020	115N04875	0.045
	115N04680	1.104	115N04339	0.012	115N11086	0.256	115N04866	0.005	115N04686	0.006
			115N04867	-2.172	115N04855	0.133	115N04186	0.003	115N04178	0.003
					115N04342	0.120			115N04179	0.001
					115N04882	0.087				
					115N04204	0.082				
					115N04672	0.065				
					115N04199	0.059				
					115N04200	0.053				
					115N04198	0.052				
					115N04867	0.048				
					115N04191	0.038				
					115N04192	0.022				
					115N07209	0.022				
					115N04681	0.007				
					115N04180	0.001				
					115N04872	-0.718				

Table 4. 7 Identified Weather Stations and AME

Weather Station	Data Type	AME	Time of Day	Direction
KAZPHOEN71	Wind Speed (MPH)	0.253	0am-6am	Westbound
KAZPHOEN209	Wind Speed (MPH)	0.048	0am-6am	Westbound
KAZBUCKE12	Temperature (F)	0.016	6am-10am	Eastbound
KAZCHAND42	Wind Speed (MPH)	0.013	10am-4pm	Eastbound



(a) Westbound



(b) Eastbound

Figure 4. 6 10am-4pm Identified Critical Segments, Weather Stations and Crashes

4.4.3 Distribution Selection

The number of I-10 truck crashes is predicted by the conditional mean, while the related probabilities are given by selected mixed Poisson distributions. To evaluate the crash risks, the prediction about individual probabilities should be provided, since it allows analysis of the tail behavior rather than only the conditional mean. First, the 7 different distributions are fitted to the data. Their common mean function is estimated through regression on buffer time of selected

locations and the remaining parameter σ and ν are determined by the Maximum Likelihood Estimator (MLE). Since conditioning σ and ν on explanatory variables hardly improves the distribution fitting results (Rigby et al., 2008) and there are no corresponding explicit interpretations, the two parameters in related distributions are treated as constants when deriving the MLEs. Notice that even though the 7 distributions have closed form of PMFs, their score equations can be very nonlinear. The numerical iterative Quasi-Newton approaches (i.e. BFGS, Nelder-Mead) are preferred in finding the MLEs. Common goodness-of-fit measures for the GLM family are the Pearson and deviance statistics. However, as it is indicated that these 2 tests are only appropriate when there are repeat observations and the number of replicates is sufficiently large (Neter et al., 1996). For the cases where the explanatory variables are continuous (the buffer time and weather data) and the number is big, the 2 conditions are seldom satisfied. Instead, we propose using the Chi-square goodness of fit test as distribution selecting measures. An intuitive view of this test is to compare fitted probabilities with real frequencies. The real frequencies are computed directly from the truck crash data sample and denoted as $\overline{p}_j$, while the fitted probabilities are calculated by $\hat{p}_j^{\,F}$, where the superscript F is indicating the kind of competing distributions.

$$\overline{p}_j = \frac{\sum_{i=1}^{n} I\left(y_i^{\,C} = j\right)}{n} \tag{4.29}$$

$$\hat{p}_j^{\,F} = \frac{\sum_{i=1}^{n} \Pr\left(\hat{y}_i^{\,C} = j\right)}{n} \tag{4.30}$$

By a comparison of these two probabilities, the competing models' tendencies to over-predict and under-predict are highlighted. Two cases are taken as examples in Figure 4.7. It is noticed that based on the same estimated mean function, different distributions do not have very different performances. To differentiate the subtle prediction difference, a formal test should be introduced. The test statistic is:

$$\sum_{j=1}^{J} \frac{\left(n\bar{p}_j - n\hat{p}_j^{\,F}\right)^2}{n\hat{p}_j^{\,F}} \tag{4.31}$$

This is a chi-square distributed statistic with $J-1$ degrees of freedom. For each of the candidate distributions, their test statistics are calculated and displayed in Table 8. The null hypothesis is that the model is correctly specified and the type I error is set 5%. The table shows the fitness performance of most mixed Poisson regressions are not significantly different. For most cases in Table 4.9, the Poisson regression, the Negative Binomial regression, the Poisson Inverse Gaussian regression and the Delaport regression achieves the best fitness result with smallest test statistics, while the Zero Inflated Poisson regression achieves the worst fitness result with the largest test statistics. Specifically, for the cases of westbound 10am-4pm, westbound 4pm-8pm and eastbound 0am-6am, the Goodness of Fit test indicates the Zero Inflated Poisson regression should be strongly rejected (test statistics greater than the Chi square statistics). For the cases of westbound 4pm-8pm, eastbound 6am-10am, and eastbound 4pm-8pm, the Sichel regression is shown to have the smallest test statistics indicating its best fitness performance. Overall speaking, on one hand, the different crash count data characteristics of Time of Day requires different kinds of mixed

Poisson distributions to account for the heterogeneity and no one regression model can achieve the best performance for all periods. On the other hand, since all the tested mixed Poisson regression models are based on the same mean function, most of their fitting performances do not diverge significantly.

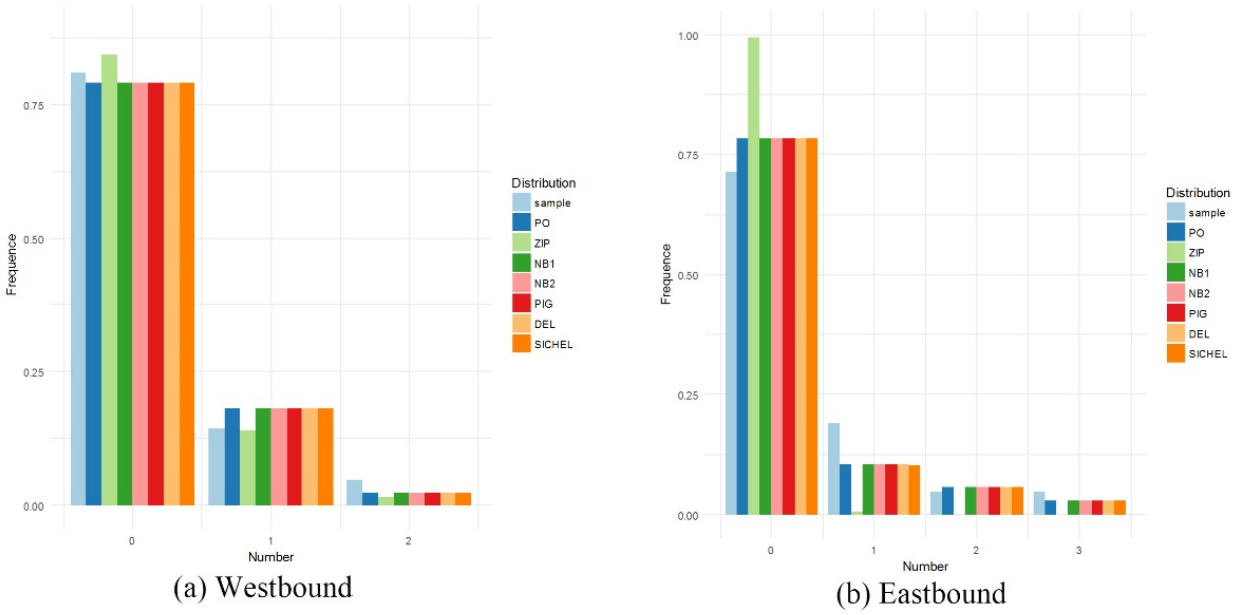


Figure 4. 7 0am-6am Candidate Distribution Goodness of Fit Histogram

Table 4. 8 Westbound Chi Square Goodness of Fit Test Statistics

	0am-6am	6am-10am	10am-4pm	4pm-8pm	8pm-12pm
PO	0.530	3.313	12.024	5.456	0.530
ZIP	1.567	3.877	34.437	88.846	1.567
NB1	0.530	3.313	12.024	5.456	0.530
NB2	0.530	3.313	12.024	5.456	0.530
PIG	0.530	3.313	12.024	5.456	0.530
DEL	0.530	3.313	12.024	5.456	0.530
SICHEL	0.537	3.338	12.024	5.513	0.537
Chi Square	5.991(df=2)	9.488(df=4)	12.592(df=6)	12.592(df=6)	3.841(df=1)

Table 4. 9 Eastbound Chi Square Goodness of Fit Test Statistics

	0am-6am	6am-10am	10am-4pm	4pm-8pm	8pm-12pm
PO	3.837	2.887	5.970	5.718	0.075
ZIP	2749760.000	3.021	7.407	6.850	0.089
NB1	3.837	2.887	5.970	1.606	0.075
NB2	3.837	2.887	5.970	1.593	0.075
PIG	3.837	2.887	5.970	1.633	0.075
DEL	3.837	2.887	5.971	1.572	0.075
SICHEL	3.874	2.823	6.043	1.589	0.077
Chi Square	7.815(df=3)	7.815(df=3)	11.071(df=5)	11.071(df=5)	3.841(df=1)

4.5 Conclusion

The chapter provides a crash frequency investigation framework relating the travel time, weather information with the crash risks. In this framework, truck crash data from Arizona I-10 is analyzed. The major results include travel time imputation, critical variable selection, crash risk assessment and crash number prediction.

Incomplete travel times are imputed using the state space model considering both the spatial and temporal factors. A DLM is introduced to describe the interdependent relationship of travel times among observation locations and epochs. The missing records are imputed using the EM algorithm based on the available travel time data. To validate the accuracy of this imputation approach, a designed experiment is conducted using censored data and raw data. The result indicates the spatial temporal imputation approach outperforms the commonly used average imputation approach.

A mean function describing expected crash numbers is developed based on all the available Time of Day segments travel time and nearby weather station information. The modern statistical variable selection method EN is implemented for critical segments and weather station identification. For most identified critical segments, their buffer times are contributing to the crash numbers. For identified weather stations, the high wind speed is shown to have the impact of increasing crash risks. These identified factors are validated using the test data set for Time of Day crash number prediction.

Based on the estimated mean function, mixed Poisson regression models with different heterogeneity distributions are introduced. Their abilities to describe crash probabilities are examined using histogram illustrations and formal tests. The test result shows that the crash risks should be quantified by different regression forms instead of one distribution with respect to the specific period of the day.

The crash number mean function and the mixed Poisson regression models together enable the ability for better crash number prediction. This framework can provide the crash numbers and expected crash probabilities given the corridor's identified segment buffer time and weather information are available. Even though the mean function can be used for estimating the expected crash incident number, the additional distribution regression model further accounts for randomness in prediction. Then, the predicted crash number is given by the mode of the corresponding best fitted PMF (the number with the biggest probability) and denoted as $\hat{y}_i^M$. Traditionally, the prediction accuracy measure is MAE or MAPE (Mean Absolute Percentage Error). The MAPE is calculated as $1/n_I \sum_{i=1}^{n_I} \left| \left(\mu_i - y_i^C \right) / y_i^C \right|$ and has the advantage of being scale-independent. However, the measure assumes that the quantity of interest is strictly positive

(Tofallis, 2015) which is not appropriate in the crash number prediction, since crash numbers being zero is not uncommon. The MAPE and its extended version Symmetric MAPE are not considered as good measures in this situation (Hyndman and Koehler, 2006). Therefore, we propose to use the Percent Better accuracy measure which is more reliable, but the price is to ignore the magnitude of the errors (Armstrong and Collopy, 1992). If the difference of predicted crash number and the real crash number within 1 is treated as better prediction results, then this measure (formula (4.32)) describes the percentage of better predictions. Except for $\hat{y}_i^M$, all other letters have same definitions with that in formula (4.27). Table 4.10 displays the prediction results. On the other hand, truck crash accidents belong to the rare events which are generally expected to happen with very low probabilities. A trivial method for predicting crash number is to assume zero accident would happen, which would also have a good prediction accuracy using our empirical knowledge. Table 4.11 shows the prediction accuracy of such empirical method. By comparing Table 4.10 and Table 4.11, we can conclude that generally the predictions by our statistical method are more accurate than that by the empirical method. Specifically, it is observed that the statistical models have the same performance with the empirical method for time period 8pm-12pm. The reason is that there are few truck accidents during that period. For the rest of time periods, the statistical models are obviously superior.

$$PB = \frac{1}{n_I}\sum_{i=1}^{n_I} I\left(\left|\hat{y}_i^M - y_i^C\right| \le 1\right) \times 100\% \tag{4.32}$$

Table 4. 10 Crash Number Prediction by Statistical Models

	0am-6am	6am-10am	10am-4pm	4pm-8pm	8pm-12pm
Westbound Training Accuracy	100.0%	92.9%	87.0%	100.0%	100.0%
Westbound Test Accuracy	94.8%	89.5%	79.0%	89.5%	94.8%
Eastbound Training Accuracy	100.0%	88.1%	88.1%	88.1%	100.0%
Eastbound Test Accuracy	94.8%	63.2%	89.5%	84.3%	100.0%

Table 4. 11 Crash Number Prediction by Empirical Knowledge

	0am-6am	6am-10am	10am-4pm	4pm-8pm	8pm-12pm
Westbound Training Accuracy	95.2%	88.1%	57.1%	83.3%	100.0%
Westbound Test Accuracy	94.8%	84.2%	42.1%	84.2%	94.8%
Eastbound Training Accuracy	90.5%	83.3%	69.1%	71.4%	100.0%
Eastbound Test Accuracy	90.8%	63.1%	47.4%	78.9%	100.0%

The approach is beneficial since the corridor's future travel times and weather information can be predicted and obtained using various methods. For example, the Kalman Filter can be used for a one-step ahead prediction for travel times based on the daily historical records. Weather forecast can be obtained from public media. Based on this information, future expected crash risks can be quantitatively assessed.

5 Vehicle Re-identification in a Connected Vehicle Environment

5.1 Introduction

Connected Vehicle (CV) technology has made the realization of Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communication possible. New applications are emerging under the V2I framework that enable the Road Side Unit (RSU) to collect real-time vehicle data to improve traffic mobility, as well as safety. According to the FHWA Vehicle-to-Infrastructure Deployment Guidance (FHWA, 2017), limited deployment of RSUs and equipped vehicles has begun and is expected to become widespread over the next 10 to 20 years. With the rapid development of CV technologies, more and more users' data will be collected.

While the enriched data facilitates CV applications, publicly publishing data could potentially pose significant privacy and security concerns. The Federal Trade Commission (FTC) and the NHTSA held an open workshop on June 28, 2017 in Washington, D.C to discuss such privacy and security issues (Juliana Gruenwald Henderson, 2017). The spatial and temporal attributes of trajectories can be used as powerful quasi-identifiers linking personal identification information (PII). Movements of connected vehicles can be observed by chance or on purpose, even with few observations. By matching trajectories having the same spatial and temporal attributes, a user' trip information may be inferred, including home and work addresses. Therefore, third party's access to such trajectory data may cause serious privacy and security threats. In the connected vehicle environment, the equipped vehicles generate random identities every 300 seconds to break one's trajectory into pieces and mix with other vehicles. However, even with such a pseudonym

protecting mechanism, exposure of the de-identified vehicle trajectory data to adversaries is still considered to be risky. This paper serves as an evidence to show that strict CV protection strategies are essential to secure consumer privacy.

When new ITS applications are implemented in the CV environment, the individual connected vehicle's travel time and delay are important measures for performance analysis. Without a vehicle ID matching process, the real-world system-level performance measures cannot be fully implemented, since one vehicle may have different IDs as it moves among intersections. Therefore, matching connected vehicles'IDs becomes desirable. More research work is needed in vehicle ID matching in the quickly maturing CV environment as the newly emerged technologies (V2I and V2V communications) are scheduled to be deployed nationwide in the next few years.

This chapter explores trajectory matching possibilities by implementing several machine learning classification techniques and systematically assessing factors potentially affecting the matching accuracy. The organization of this chapter is: Section 5.2 provides the description of connected vehicle re-identification. Section 5.3 introduces machine learning methods for CV matching. Section 5.4 provides results including model specification, validation and factor effect analysis. Section 5.5 summarize the chapter.

5.2 Description of Connected Vehicle Re-identification

Typically, RSUs are mounted on top of traffic light signal pole of equipped intersections. Using Dedicated Short Range Communication (DSRC), approaching vehicles send BSMs up to 10 times per second to an RSU. According to the Multi-Modal Intelligent Traffic Signal System (MMITTS) project (University of Arizona, 2016), for purposes of traffic signal control and performance

measures, BSMs are collected and stored every second when vehicles are in the range of DSRC which is between 300 and 600 meters. More details of this database design can be found on Page 61 of the final report (University of Arizona, 2016).

According to the Society of Automotive Engineers (SAE) J2735 standard (Kenney, 2011), the BSM data contains vehicle position, time, speed, heading acceleration, brake system status, phases, etc. In this research, only the vehicle position and the timestamp are used to match CV IDs and other vehicle properties are not included in the matching algorithms. This is due to two considerations: First, vehicle physical properties can serve as the quasi-ID to match vehicles before the machine learning techniques are applied. This pre-process can reduce the number of vehicle to be matched. In the situation where each vehicle has its unique properties, the reidentification can be done without further matching algorithms. But the methods proposed in this paper can deal with more complex scenarios in which many vehicles share common physical properties. Second, incorporating the physical properties as covariates will cause computation difficulties. For example, the constant values will make covariance matrices singular for the LDA and QDA methods.

5.3 Machine Learning Methods

Trajectory attributes used to train classifiers are a vehicle's location coordinates and corresponding time stamps. For a pair of intersections (A and B in Figure 5.1), the training dataset is from one intersection RSU's BSM data and the test dataset is from the other intersection. Let Y represent the ID variable and a 3×1 vector variable $\mathbf{X}=\begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix}^T$ represent the trajectory attributes in which X_1 and X_2 are vehicles' local coordinates with the origin defined by the intersection

center's coordinates (Barth and Farrell, 1999). X_3 represents the trajectory's time stamp. With a trained classifier, the identity of a vehicle is determined by its trajectory points' most classified ID. In this research, an odd number of trajectory points (i.e. 5 trajectory points) is used when the trained model is applied to the test dataset.

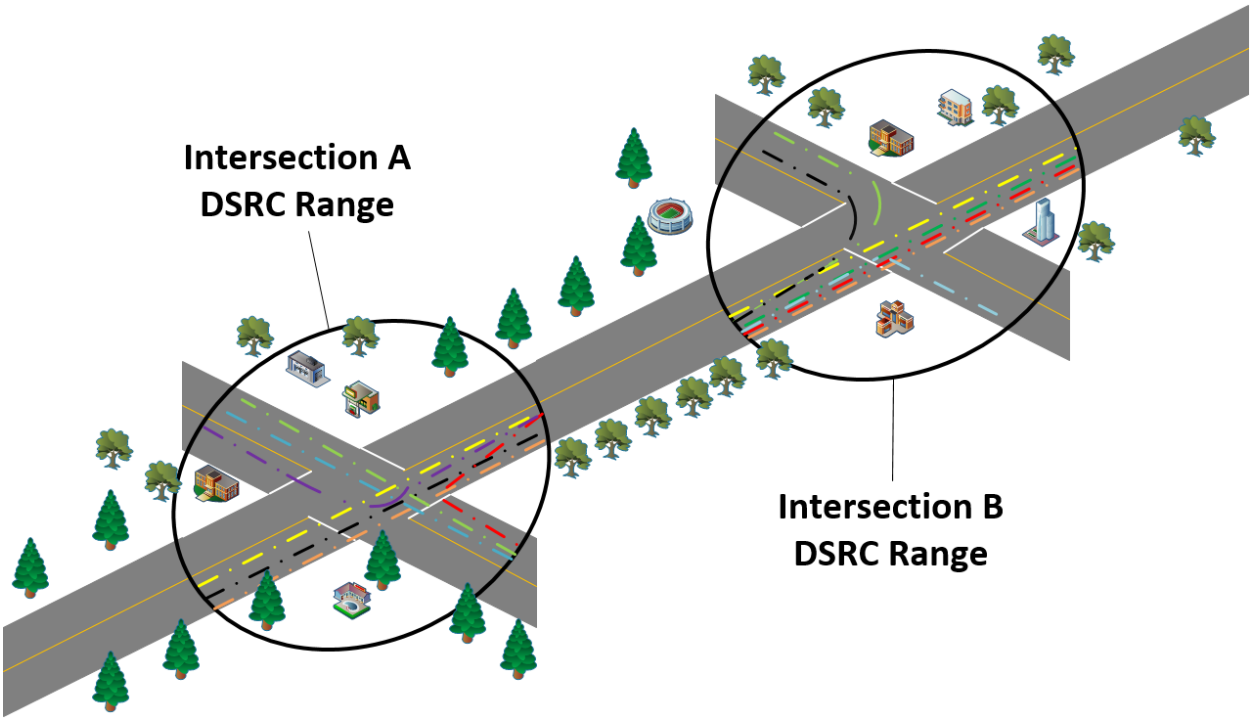


Figure 5. 1 Vehicle Trajectories

5.3.1 Logistic Regression

LR models the boundary between two trajectory attributes using linear functions with the logit transformation:

$$\log \frac{\Pr\left(Y = i \mid \mathbf{X} = \boldsymbol{x}\right)}{\Pr\left(Y = K \mid \mathbf{X} = \boldsymbol{x}\right)} = \beta_{i0} + \boldsymbol{\beta}_i^T \boldsymbol{x} \tag{5.1}$$

In which $i = 1, \cdots, K-1$. K represents the order of ID in the sample. Denote the parameter vector $\boldsymbol{\theta} = \left\{\beta_{10}, \boldsymbol{\beta}_1, \cdots, \beta_{(K-1)0}, \boldsymbol{\beta}_{K-1}\right\}$. Given a trajectory attribute, the probability of being ID i is:

$$p_i\left(\boldsymbol{x}\right) = \Pr\left(Y = i \mid \boldsymbol{x}\right) = \begin{cases} \dfrac{\exp\left(\beta_{i0} + \boldsymbol{\beta}_i^T \boldsymbol{x}\right)}{1 + \sum_{j=1}^{K-1} \exp\left(\beta_{j0} + \boldsymbol{\beta}_j^T \boldsymbol{x}\right)} & i = 1, \cdots K-1 \\ \dfrac{1}{1 + \sum_{j=1}^{K-1} \exp\left(\beta_{j0} + \boldsymbol{\beta}_j^T \boldsymbol{x}\right)} & i = K \end{cases} \tag{5.2}$$

The parameter estimation is based on MLE and the likelihood is $l\left(\boldsymbol{\theta}\right) = \sum_{v=1}^{n} \log p_{y_v}\left(\boldsymbol{\theta}; \boldsymbol{x}_v\right)$, in which n is the total number of trajectory attributes in the training set. Since the likelihood is of a complex form, it can be solved by the traditional numerical method Newton-Raphson algorithm (Rizzo, 2007). Given an unknown vehicle's trajectory attribute $\boldsymbol{x}$, the assigned ID is:

$$ID = \arg\max_i p_i\left(\boldsymbol{x}\right) \tag{5.3}$$

5.3.2 Linear Discriminant Analysis

The LDA is another popular linear classification model. The assumption is that each trajectory's composing point density is multivariate Normal $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ with the same covariance matrix $\boldsymbol{\Sigma}$. Let π_i be the prior probability of trajectory i $(i = 1, \cdots, K)$ and $g_i(\boldsymbol{x})$ be the *ith* trajectory-conditional density. Then the posterior ID probability is $\Pr(Y = i \mid \boldsymbol{x}) = g_i(\boldsymbol{x})\pi_i \Big/ \sum_{j=1}^{K} g_j(\boldsymbol{x})\pi_j$. The discriminant function (Equation (5.4)) for each trajectory ID is obtained using $\log\{\Pr(Y = i \mid \boldsymbol{x}) / \Pr(Y = j \mid \boldsymbol{x})\}$ and some simple algebra.

$$f_i(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \log \pi_i \tag{5.4}$$

The parameter estimation is defined by Equation (5.5-5.8)

$$\hat{\pi}_i = \frac{n_i}{n} \tag{5.5}$$

$$\hat{\boldsymbol{\mu}}_i = \sum_{i=1}^{K} \frac{\boldsymbol{x}_i}{n_i} \tag{5.6}$$

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{K} \frac{n_i - 1}{\sum_{j=1}^{K} (n_j - 1)} S_i \tag{5.7}$$

$$S_i = \frac{1}{n_i - 1} \sum_{Y_v = i} (\boldsymbol{x}_v - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_v - \hat{\boldsymbol{\mu}}_i)^T \tag{5.8}$$

In which S_i is *ith* trajectory attributes' covariance matrix. Given an unknown vehicle's trajectory attribute $\boldsymbol{x}$, the assigned ID is:

$$ID = \arg\max_i f_i(\boldsymbol{x}) \tag{5.9}$$

5.3.3 Quadratic Discriminant Analysis

The QDA model assumptions are similar to LDA's and the only difference is that the trajectories' multivariate Normal $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ densities can have their own covariance matrix $\boldsymbol{\Sigma}_i$. The corresponding discriminant function is:

$$f_i(\boldsymbol{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \log\pi_i \tag{5.10}$$

The parameter estimation procedure (Equation (5.5-5.7)) is applied to QDA, but the covariance matrix estimation is different:

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1}\sum_{Y_v = i}(\boldsymbol{x}_v - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_v - \hat{\boldsymbol{\mu}}_i)^T \tag{5.11}$$

5.3.4 Linear Support Vector Machine

In the Linear SVM literature, the boundary separating two trajectories is also in the linear form. The decision function is of the form $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$ and the parameter estimation is based on the idea of maximizing the margin between the two trajectories (Vapnik, 2013). Solving such a problem is an exercise in the convex optimization. Given that the two trajectories' IDs are coded as -1 and +1, the popular setup is:

$$\min_{\beta_0, \boldsymbol{\beta}} \ \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N} \xi_i \tag{5.12}$$

$$\text{subject to} \quad \xi_i \geq 0, y_i\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0\right) \geq 1 - \xi_i, \forall i \tag{5.13}$$

In which ξ_i is the non-negative slack variable allowing trajectory attributes to be on wrong side of the decision boundary. N is the total number of two training trajectories' attributes. C is a cost parameter that controls the overlap and is decided by a 10-fold Cross Validation (Geisser, 1975). A quadratic programming solution of $\boldsymbol{\beta}$ and β_0 using Lagrange multipliers was introduced by (Friedman et al., 2001). The solution for $\boldsymbol{\beta}$ has the form $\hat{\boldsymbol{\beta}} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \boldsymbol{x}_i$ in which $\hat{\alpha}_i$ is a positive Lagrange multiplier and the resulting decision function is expressed by Equation (5.14) in which $\langle \cdot, \cdot \rangle$ is the inner product operator. Given an unknown vehicle's trajectory attribute $\boldsymbol{x}$, the assigned ID is indicated by Equation (5.15).

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle \tag{5.14}$$

$$ID = sign\left[\hat{f}(\boldsymbol{x}) \right] \tag{5.15}$$

For $m \geq 2$ vehicle ID matching problems, a 'one against one' approach is adopted and $m(m-1)/2$ binary decision functions are estimated. The final decision is made based on the most identified ID by all the binary classifiers.

5.3.5 Nonlinear Support Vector Machine

The nonlinear SVM allows more general decision surfaces than linear boundaries of the linear SVM. This flexibility comes from the enlarged feature space consisting of the expanded basis function $h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \cdots, h_M(\boldsymbol{x}))$. The corresponding estimated decision function is indicated by Equation (5.16).

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle h(\boldsymbol{x}), h(\boldsymbol{x}_i) \rangle \tag{5.16}$$

In which the inner product of expanded basis $\langle h(\boldsymbol{x}), h(\boldsymbol{x}_i) \rangle$ is further defined by the kernel function $K(\boldsymbol{x}, \boldsymbol{x}_i)$. Two popular choices for the K function are implemented in this research, the polynomial kernel function by Equation (5.17) and the Gaussian kernel function by Equation

(5.18). The parameter d and σ are the tuning parameters determined using a 10-fold Cross Validation. Finally, the ID matching rule is same as formula (5.14).

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \left(1 + \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle\right)^d \tag{5.17}$$

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 / \sigma\right) \tag{5.18}$$

5.3.6 K Nearest Neighbor

The KNN classifier needs no model to be fit. The idea of this matching mechanism is to assign the ID to vehicle trajectory attribute based on the closest K training examples in the spatial and temporal space. The attribute $\boldsymbol{x}$ is assigned to the ID most common amongst its K nearest neighbors. The metric to measure distances between attributes is the Euclidean distance. The tuning parameter K is decided by a 10-fold Cross Validation.

5.4 Results

5.4.1 Experiment Design

To investigate the matching accuracy, a one-hour simulation data was collected. Reasons for using simulation data are: first, there are only a few OBUs (less than 10) in deployment. The small quantity of OBUs limits the experimental capability to investigate basic scenario impacts on matching accuracy. Second, using simulation data renders us able to collect BSM data across all

kinds of condition combinations, some of which are rarely seen in real world (e.g. 100% CV market penetration rate)

Besides the method factor, other factors include the traffic volume, CV market penetration rate, number of vehicles to be matched and the location. Among those four factors, the traffic volume, CV market penetration and number of vehicles to be matched are treated as random factors, indicating that if the experiment is conducted again, the treatment levels may be different. Three traffic volume levels are selected randomly: 200 veh/h on the main street and 100 veh/h on side streets for the low volume, 400 veh/h on the main street and 200 veh/h on side streets for the medium volume, 800 veh/h on the main street and 400 veh/h on side street for the high volume. Different CV market penetration rates are: 20%, 60% and 100%. Six numbers of vehicles are matched: 2, 5, 10, 30, 60 and 100. The method and the location are regarded as fixed factors. Four pairs of intersections are selected from the MMITTS CV test corridor at Anthem, Arizona (Figure 5.2).

Figure 5. 2 CV Test Corridor

Due to the traffic volume limitations, the number of vehicles to be matched and the traffic volume levels cannot be fully crossed in this experiment. For the low traffic volume case, only 2, 5, 10 and 30 vehicles are used for ID matching. For the medium traffic volume, additional 60 vehicles are used. For the high traffic volume case, all the vehicle number levels are used. For some vehicle number and traffic volume combinations, there is no result data collected. This characteristic makes this experiment an unbalanced one. Vehicles' trajectory data is collected through the SIL technique in which the RSU applications collect BSMs sent by CVs in the VISSIM simulation environment (University of Arizona, 2016). For each combination of the four factors, five

replications of matching experiments are conducted by different random seeds in the R software environment.

The total number of matching records is 6300 and the average mis-matching rates are summarized by factor in Table 5.1. A first look at the table reveals that the Logistic Regression has the worst performance among all test methods. The LDA, QDA and Linear SVM have the relatively smallest mis-matching rates. More vehicles lead to higher mis-matching rates. Differences of mis-matching rates exist among subcategories of Location, Market Penetration Rate and Traffic Volume. Due to the complexity and unbalanced characteristics of this experiment, a formal analysis is proposed below.

Table 5. 1 Average Mis-Matching Rates

Method	Average	Numer of Vehicles to Be Mateched	Average	Location	Average	Market Penetration Rate	Average	Traffic Volume	Average
LR	0.82	2 Veh	0.10	Intersection Pair 1	0.17	20%	0.30	Low	0.23
LDA	0.14	5 Veh	0.17	Intersection Pair 2	0.44	60%	0.31	Medium	0.30
QDA	0.14	10 Veh	0.26	Intersection Pair 3	0.27	100%	0.31	High	0.36
Linear SVM	0.14	30 Veh	0.41	Intersection Pair 4	0.33				
Guassian Kernel SVM	0.45	60 Veh	0.53						
Polynomial Kernel SVM	0.22	100 Veh	0.63						
KNN	0.21								

Since the mis-matched number is a non-negative integer, for the purpose of investigating impacts of associated fixed and random factors, a Poisson mixed regression model is appropriate. The traditional standard Poisson regression usually suffers from the heterogeneity problem in count data. Over dispersion is the major heterogeneity form in Poisson distribution (Cameron and Trivedi, 2013). In other words, only the fixed and random factors cannot guarantee to explain all

the variations in the data. This extra variation occurred in modeling count data is accounted by an additional random residual term. Equation (5.19) and (5.20) are proposed for modeling relations between the count response and the latent variable, which is further predicted by a linear model.

$$z_i \sim Poi(\lambda_i = \exp(l_i)) \tag{5.19}$$

$$l_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{u}_i^T \boldsymbol{\eta} + e_i \tag{5.20}$$

Where z_i represents the ith record of mis-matched number, following the Poisson distribution with the rate λ_i. In the Generalized Linear Model family, the rate λ_i is related to the latent variable l_i given that a canonical log link function is employed (Nelder and Baker, 1972b). Here variables $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are reused. $\boldsymbol{x}_i$ is a 10×1 vector $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, ..., x_{i9})^T$ representing the intercept and fixed factor indicator variables and $\boldsymbol{\beta}$ is a 10×1 vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_9)^T$ representing the corresponding effects. $\boldsymbol{u}_i$ is a 3×1 vector $\boldsymbol{u}_i = (u_{i1}, u_{i2}, u_{i3})^T$ representing 3 random factor indicator variables and $\boldsymbol{\eta}$ is a a 3×1 vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^T$ representing the corresponding effects. e_i is the residual term.

The MCMC sampling technique within the Bayesian hierarchical model framework is used for exploring both the fixed and random factors' effects. We follow the work of (Hadfield, 2010) and make multivariate normal distribution assumption (Equation (5.21)) for the effect and residual parameters ($\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\boldsymbol{e} = (e_1, \cdots, e_{6300})^T$).

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\eta} \\ e \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{B} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{R} \end{bmatrix} \right) \tag{5.21}$$

In which $\boldsymbol{\mu}$ is the prior mean for fixed effects with prior covariance matrix $\boldsymbol{B}$. For random factors and residual term, their prior means are set to 0 and covariances are $\boldsymbol{G}$ and $\boldsymbol{R}$. Parameters can be Gibbs sampled and details can be found in (García-Cortés and Sorensen, 2001; Hadfield, 2010). Before proceeding to the sampling process, parameter prior distributions should be set up first. For fixed factor, a normal prior is set with $\boldsymbol{\mu} = (0, \cdots, 0)^T$ and $\boldsymbol{B} = diag\left\{10^8, \cdots, 10^8\right\}$. For random and residual variance components ($\boldsymbol{G} = diag\left\{G_1, G_2, G_3\right\}$ and $\boldsymbol{R} = diag\left\{r, \cdots, r\right\}$), inverse gamma distributions are assigned with shape and scale parameters set to 0.001. The MCMC sample size is 6×10^6 and the first 6×10^5 samples are discarded as the burn-in process. Each parameter sample is collected every 500 iterations for the purpose of being identically and independently distributed.

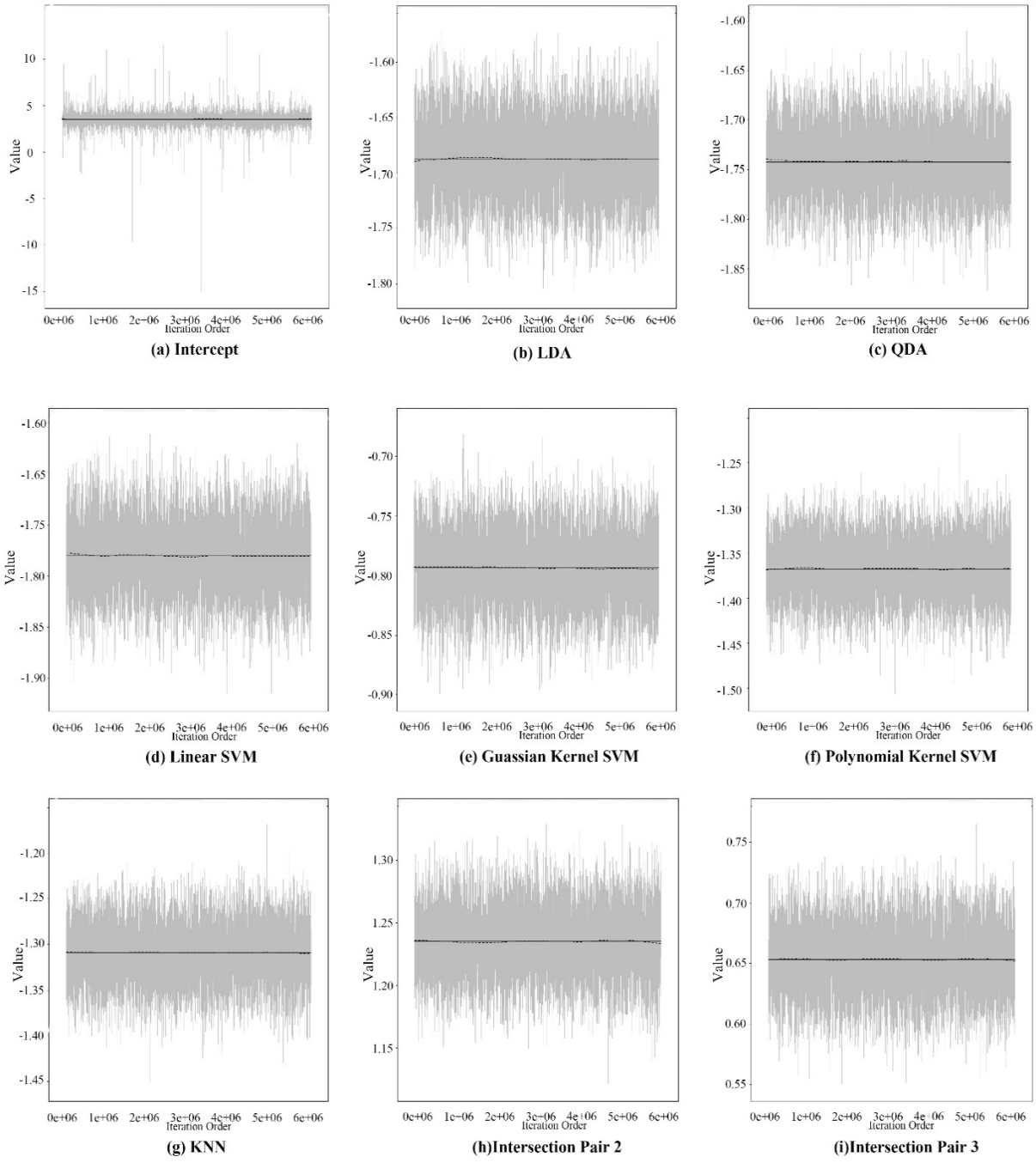
5.4.2 Model Validation

The purpose of the Bayesian model validation is to check whether the parameter samples drawn by the MCMC can accurately represent their true posterior distributions.

Visualized Convergence Check

Parameter Samples of posterior distributions are plotted in Figure 5.3(a)-5.3(n). The constant variation and the unchanging mean are good signs of convergence of the sampling process. All the fixed effect and the residual parameters perform well, but the random effect parameter chains (Figure 5.3(k), 5.3(l) and 5.3(m)) for the variances are spiky due to their posterior distributions having long right tails. It is helpful to plot them on a log scale to reduce the influence of a few

outliers. In Figure 5.3(o-q), the parameters are behaving better. The visual check implies good convergence of posterior distributions. Additionally, the autocorrelation check (Cryer and Chan, 2008) is used and shows the evidence of samples being independently distributed.



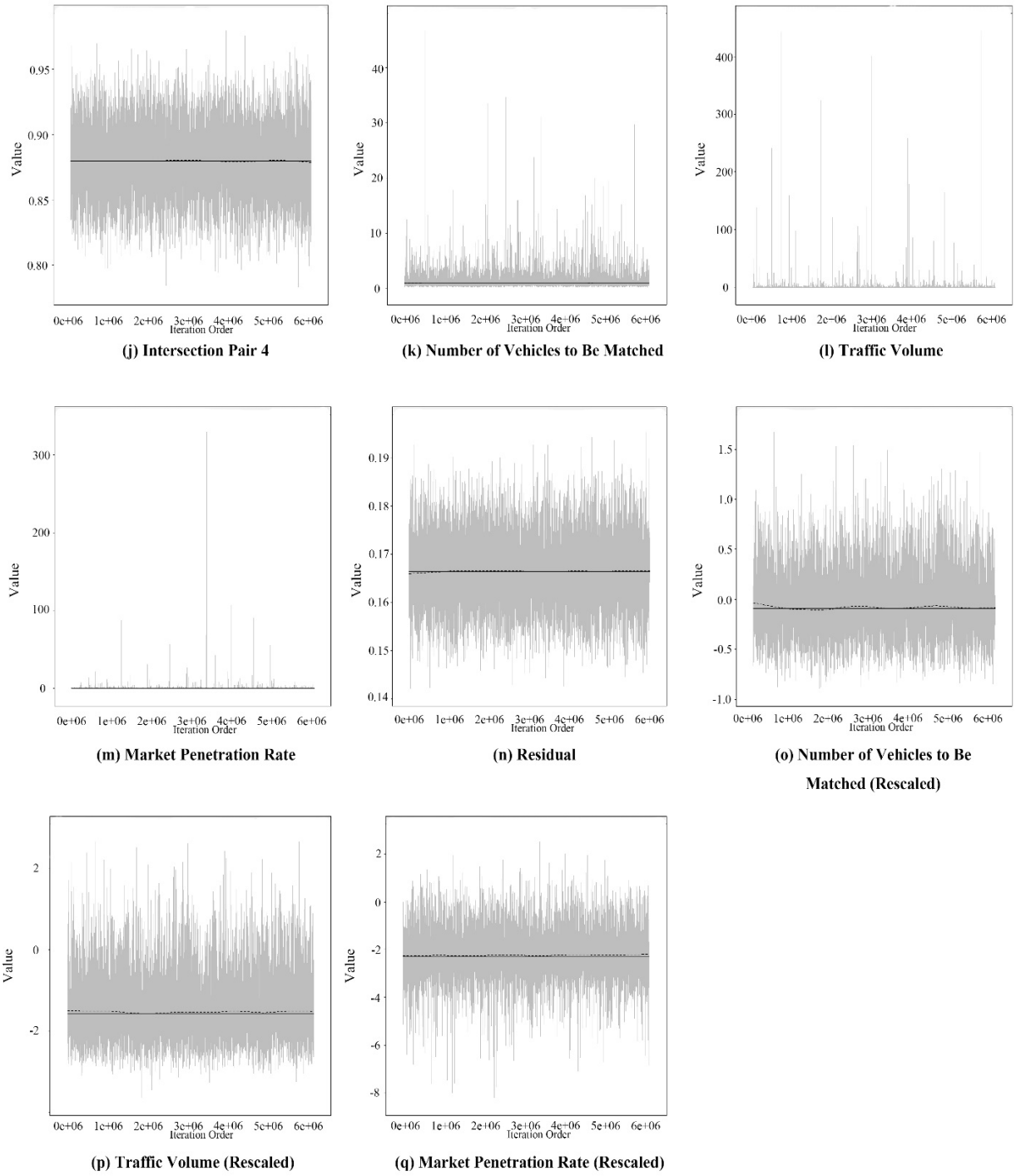


Figure 5. 3 Time Series of Parameter Traces

Gelman-Rubin Test

The stricter validation approach relies on test statistics and sampling properties. The Gelman-Rubin test (Gelman and Rubin, 1992) is used. Several independent MCMC experiments are checked whether they converge to the same posterior distribution. By comparing estimated between-chain and with-in chain variances, the sampling chain's convergence is examined. Assuming there are M chains and each has a length of L. Let $\hat{\gamma}_i$ be the sampled parameter value mean and $\hat{\sigma}_i^2$ be the sample variance of ith chain. Let $\hat{\gamma} = (1/M)\sum_{i=1}^{M}\hat{\gamma}_i$ be the overall sample posterior mean. Then the between-chain variance is defined by $V_B = \left[L/(M-1)\right]\sum_{i=1}^{M}(\hat{\gamma}_i - \hat{\gamma})^2$ and the within-chain variance is defined by $V_W = (1/M)\sum_{i=1}^{M}\hat{\sigma}_i^2$. An unbiased estimator of the marginal posterior variance of γ is defined by $\hat{V} = \left[(L-1)/L\right]V_W + \left[(M+1)/ML\right]V_B$. In (Brooks and Gelman, 1998), the Potential Scale Reduction Factor (PSRF) test statistic is defined by Equation (5.22)

$$R_c = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\frac{\hat{V}}{V_W}} \tag{5.22}$$

In which $\hat{d}$ is an estimate of degree of freedom of a *t* distribution. If the M chains have converged to the target posterior distribution, then the value of R_c should be close to 1. Practically, if $R_c < 1.2$ for all parameters, it is safe to say the convergence has been reached. An upper limit (Equation 5.23) is defined for monitoring the test statistic (Gelman and Rubin, 1992).

$$R_u(\alpha) = \sqrt{\frac{\hat{d}+3}{\hat{d}+1}\left(\frac{L-1}{L}V_W + \frac{M+1}{M}q_{1-\alpha/2}\right)} \tag{5.23}$$

In which $q_{1-\alpha/2}$ is the $(1-\alpha/2)$ quartile of an F distribution. By calculating the PSRF as the function of sampled parameter values, one can tell when the test statistic is really converged and not just randomly near 1. Figure 5.4 shows a plot ($M = 4, L = 6\times10^6$) of PSRFs (black solid lines) with the upper limits (red dashed lines) at a 95% confidence level. Note that the Gelman-Rubin convergence test assumes normality for the parameter's marginal posterior distribution. But for the Poisson regression model, there is no available conjugate normal posterior distribution. Therefore, those test statistics are calculated based on data normalized by the logarithm transformation. All parameters' test statistics are approaching 1 quickly, indicating the convergence is achieved. Therefore, the Poisson mixed model parameter samples drawn by the MCMC are validated and are treated as posterior distributions for the further effect analysis.

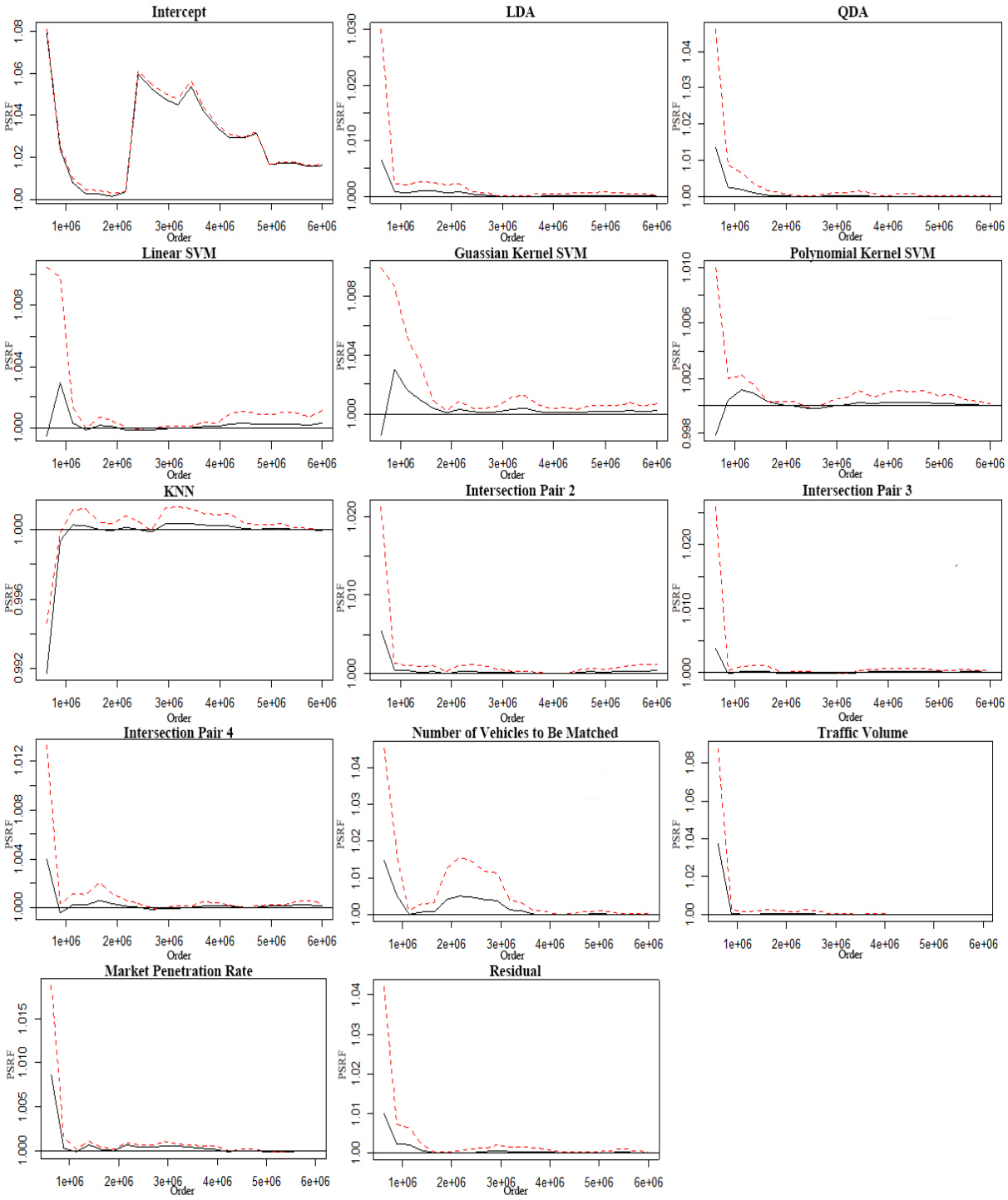


Figure 5. 4 PSRF Plot

5.4.3 Factor Effect Analysis

After obtaining effect parameters' sampled distributions, their contribution to the mis-matching vehicle rate (per 100 vehicles) should be examined. Medians are used as the estimators. The Highest Posterior Density interval is reported as the Bayesian equivalent Confidence Interval (CI). Table 5.2 shows the parameter medians, upper and lower limits with 95% CI and the effective sample sizes. Effective sample sizes are different due to the MCMC sampling rejection rule (Rizzo, 2007), if no sample is rejected then the sample size is $\left(6\times10^{6}-6\times10^{5}\right)\big/500=10800$.

Table 5. 2 Posterior Distribution Summary

Factor Type	Factor	Factor Level	Sample Median	Lower 95% CI	Upper 95% CI	Effective Sample Size
Fixed	Intercept	Intersection Pair 1+LR	3.55	2.33	4.67	10513
	Method	LDA	-1.69	-1.75	-1.62	10800
		QDA	-1.74	-1.80	-1.68	10800
		Linear SVM	-1.78	-1.85	-1.71	9507
		Gussian Kernel SVM	-0.79	-0.85	-0.74	10800
		Polynomial Kernel SVM	-1.37	-1.43	-1.31	9014
		KNN	-1.31	-1.37	-1.25	10334
	Location	Intersection Pair 2	1.24	1.18	1.28	10800
		Intersection Pair 3	0.65	0.60	0.71	10800
		Intersection Pair 4	0.88	0.83	0.93	10800
Random	Number of Vehicles to Be Matched	NA	0.81	0.15	3.34	6573
	Traffic Volume	NA	0.03	0.00	1.21	10800
	Market Penetration Rate	NA	0.01	0.00	0.34	10800
	Residual	NA	0.17	0.15	0.18	10800

The factor effects on mis-matching rates differ among locations and methods. All fixed factors' CIs do not include 0, indicating their impacts are significant. For the default level combination (the Intercept), location Intersection Pair 1 and LR, the mean of the expected mis-matching rate is 35 per 100 vehicles ($\exp\left(3.55\right)\approx35$). Given the matching is conducted in the same intersection pair,

all the other matching methods have better performance than LR, since their coefficients' sample medians have negative impacts on the mis-matching rate. Specifically, the Linear SVM ($\exp(3.55-1.78)\approx 5.87$), the QDA ($\exp(3.55-1.74)\approx 6.11$) and the LDA ($\exp(3.55-1.69)\approx 6.42$) have the best top 3 accuracy performances. All the other locations have worse performance than Intersection Pair 1, since their medians all have positive impacts on the mis-matching rate. The performance ranking is consistent with their within distances. For example, Intersection Pair 1 has the shortest within distance and the smallest mis-matching rate if the method is kept unchanged, while the Intersection Pair 2 has the largest within distance and the largest mis-matching rate (sample median value 1.24).

For the 3 random factors, their variances on the mis-matching rate are the experiment designer's interest (Montgomery, 2017), and therefore their specific levels' impacts are not included in the model. The variance components of the traffic volume factor and the market penetration rate factor have the least impacts on the mis-matching rate, since their sample medians are 0.03 and 0.01. This is suggesting that if the ID matching is conducted in a different traffic volume situation or in a different CV market penetration rate scenario, given other factors are kept unchanged, there will be little difference in the result's accuracy. To the contrary, the variance component of the number of vehicles to be matched factor has a large influence on the mis-matching rate with a median value 0.81. This is suggesting that conducting the CV ID matching with 10 vehicles will have significant difference in the result's accuracy compared with doing that with 60 vehicles or 100 vehicles. Meanwhile, the residual's variance component is not negligible (sample median value 0.17), indicating that the data is not strictly Poisson distributed but with slight over dispersion.

5.5 Conclusion

The advantages of proposed methods include 3 aspects: First, this chapter demonstrates how to use the structured CV data rather than the unstructured image data for vehicle reidentification. Seven popular machine learning classification methods are implemented using the vehicle's trajectory attributes. Second, potential factors impacting matching accuracy are systematically investigated. A Poisson mixed regression model is proposed to analyze the effects of potential factors on matching accuracy. The analysis is conducted based on the Bayesian inference framework. The experiment result shows that the Linear SVM, the QDA and the LDA are the 3 best techniques out of all the candidate machine learning methods discussed in this paper for CV ID matching. Additionally, the location factor and the number of vehicles to be matched factor have significant impacts on matching accuracy, while the traffic volume factor and the CV market penetration factor are not considered significant. Third, this research work can help with future CV based research and also serve as an evidence for public concern. Note that the smallest average mis-matching rate is 14% achieved by the 3 mentioned machine learning techniques (Table 5.1). This possibility of using trajectory data to reidentify connected vehicle ID with high accuracy renders researchers capable of assessing performance of CV based applications in the situation where vehicle IDs are changing frequently. However, securer protection mechanisms for trajectory data should be investigated and the access to trajectory data should be carefully controlled.

6 Vehicle Trajectory Prediction in a Connected Vehicle Environment

6.1 Introduction

Connected Vehicle (CV) technology is quickly becoming popular as more and more field tests are being conducted and presented. One of the most distinguished characteristic of this framework is that it can provides Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communication in real time. This advantage is enabling researchers and companies to develop methods to improve traffic mobility and safety based on the data collected through above technologies. With more deployment on its way, there will be definitely more and more such data available.

Many vehicle conflicts and crashes can be avoided if the drivers are warned potential dangers ahead. Vehicle trajectory prediction serves the purposes of providing such important information to drivers or vehicle itself (in an automated vehicle setting). While a vehicle without V2V and V2I technologies can hardly detect potential collision in its surroundings, the CV environment is ideal for implementing continuous vehicle trajectory prediction tasks. In a CV environment, a connected vehicle can communicate with its surrounding vehicles and infrastructures to exchange data and information. Using proper data aggregation and advanced prediction algorithms, it is possible to reduce the crash risks, saving lives and properties.

This chapter is organized as follows. The following section summarizes previous research work related to this study. Section 6.2 introduces the CV data used in this research and potential application scenarios. Section 6.3 presents the modeling methods, including the baseline model,

deep learning based models and advanced feature engineering techniques. Experiment design and error analysis are provided in Section 6.4. Section 6.5 summarizes this chapter.

6.2 Data Overview and Potential Application Scenarios

6.2.1 Basic Safety Message

V2V and V2I communications rely on Dedicated Short Range Communication (DSRC) standard. The Federal Communication Commission (FCC) assigned 5.9GHz bandwidth to Intelligent Transportation Systems (ITS) vehicle safety and mobility applications. The Society of Automotive Engineers (SAE) has published the Dedicated Short Range Communications (DSRC) Message Set Dictionary, J2735, (SAE International, 2009) which defines the messages for both safety and mobility applications. Message types include the Basic Safety Message (BSM), Signal Phasing and Timing Message (SPaT), MAP Message (MAP) and so on.

The BSM reports the current vehicle's location and states, including vehicle's temporary identification, position (Latitude, Longitude, and Elevation), Motion (Speed, Heading, Steering Wheel Angle and 4-way Accelaration), Control (Break System Status), etc. The specified transmission rate can be as high as 10 times per second and can be adjusted according to different needs.

6.2.2 Signal Phase and Timing Message

The Signal Phase and Timing (SPAT) message reports the intersection's current status of traffic signals. Combined with intersection's specific information, the message describes current signal

status (green, yellow, red), remaining phase time, and next phase. In addition, current signal preemption and priority status are included. In a typical signal actuation control strategy, each phase's green time can be adjusted according to lane detectors, therefore each phase's green time can be flexible. The principle of the commonly implemented vehicle extension call is illustrated in Figure 6.1. The original minimal green time will add extension on itself when a vehicle approaches in the corresponding phase. An extension interval varies from 0 second to 5 seconds according to demands. The newly updated green time length can be as long as the pre-defined maximal green time.

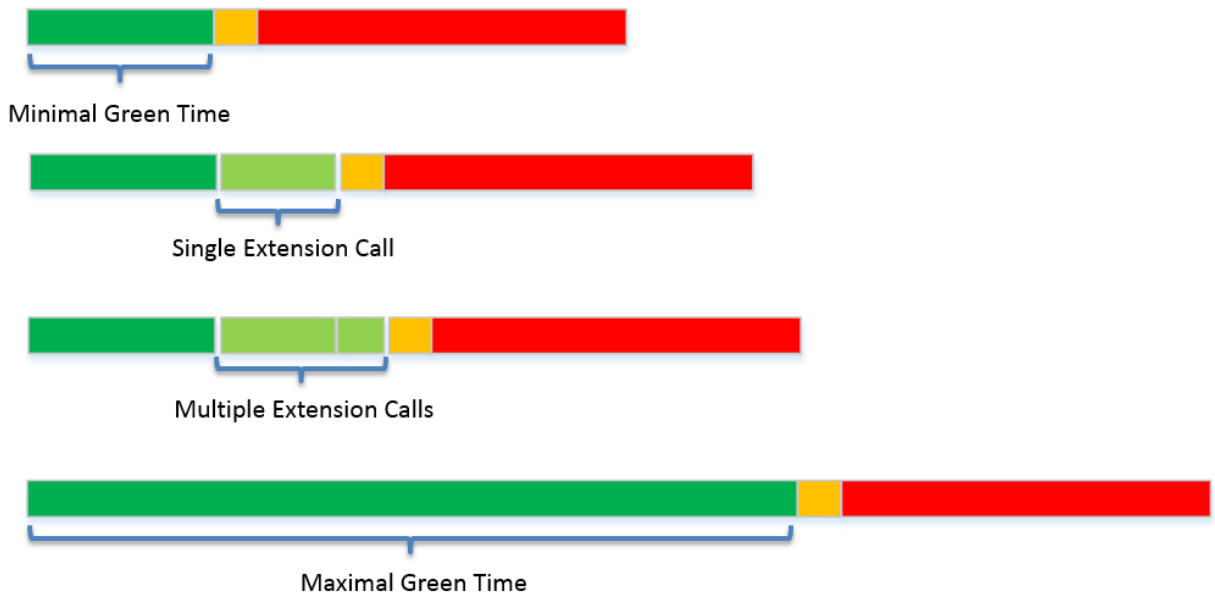


Figure 6. 1 Vehicle Extension Call

6.2.3 Potential Safety Application Scenarios

The trajectory prediction can be used in a Collision Avoid System (CAS). A common conflict scenario is the “left turn” conflict shown in Figure 6.2. In a permissive left turn setting, the driver without the right of way may run into the vehicle in the conflict phase due to falsely judging the current condition. Such conflicts may result in very severe crash accidents on road.

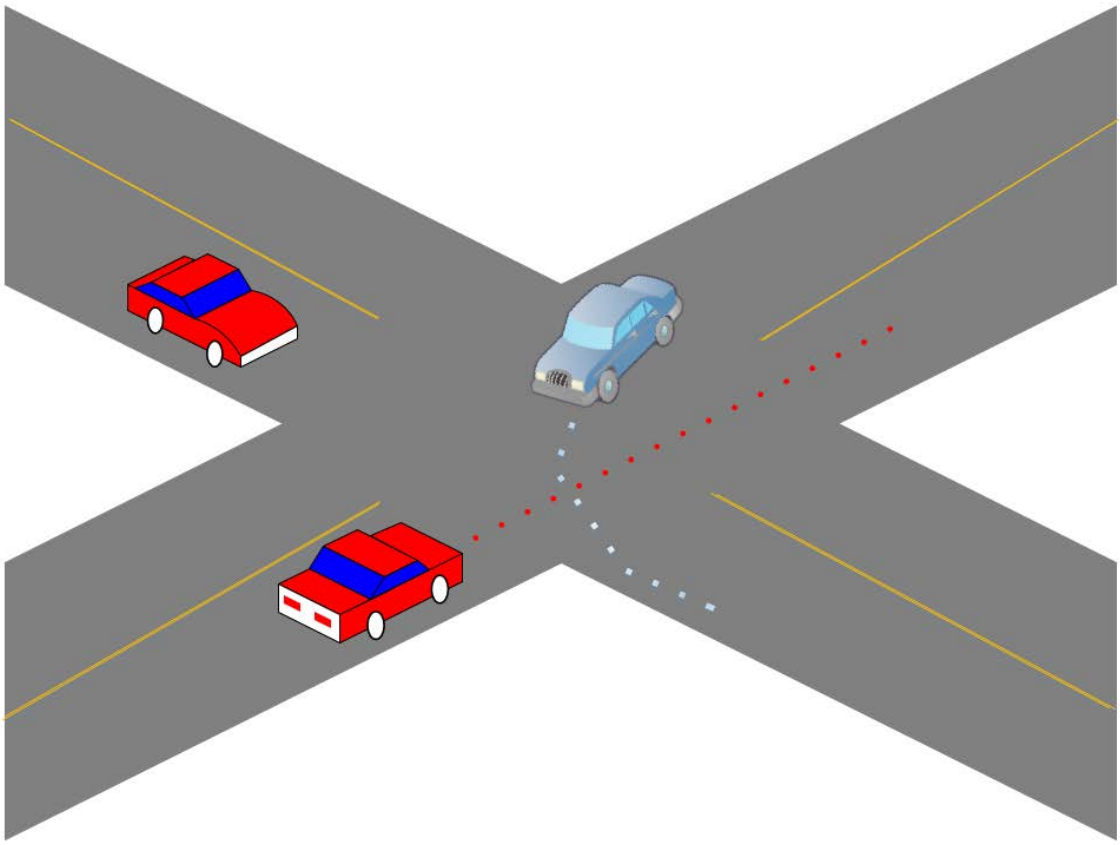


Figure 6. 2 Left Turn Conflict

6.3 Methodology

6.3.1 Baseline Model

The baseline model chosen for this research is the Kalman Filter's (KF) one step ahead prediction model. The KF model and its variants has been used widely in robotics control and moving object trajectory tracking for a very long time. A carefully tuned KF model may achieve good prediction in various application scenarios. The Model consists of two parts, the state equation and the observation equation.

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t \tag{6.1}$$

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t \tag{6.2}$$

where $\boldsymbol{x}_t$ is the $p \times 1$ state vector at time $t = 1,2,\cdots,n$ and it contains 4 scalar variables representing vehicle's state at time t, including speed, acceleration, local horizontal coordinate and local vertical coordinate. It assumes that the process start with a normal vector $\boldsymbol{x}_0 \sim N_4(\boldsymbol{\mu}_0, \Sigma_0)$. The 4×4 transition matrix Φ is indicating the spatial and temporal interaction relations in travel time evolutions. The randomness of the state process is represented by the 4×1 vector $\boldsymbol{w}_t \sim N_4(\boldsymbol{0}, Q)$. In the observation equation, $\boldsymbol{y}_t$ is the 4×1 direct observed BSM data at time t and A_t is the 4×4 measurement matrix. The data is assumed to be observed with additional noise $\boldsymbol{v}_t \sim N_4(\boldsymbol{0}, R)$. Given the formulation of the state space model, vehicle's expected positions can be predicted according to previous prediction and adjustment. The details of how Kalman Filter works can be found in Appendix B. Parameter estimation is conducted using Maximum Likelihood Estimation.

Let $\boldsymbol{x}_t^{t-1} = E\left(\boldsymbol{x}_t \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{t-1}\right)$ and $P_t^{t-1} = E\left\{\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}\right)^T\right\}$. The likelihood is computed using the innovation term $\boldsymbol{\varepsilon}_t = \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1}$. It has a mean vector of zeros and the covariance matrix is $\Sigma_t = A_t P_t^{t-1} A_t^T + R$. The log likelihood can be expressed as:

$$L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log\left|\Sigma_t\right| - \frac{1}{2}\sum_{t=1}^{n}\boldsymbol{\varepsilon}_t^T \Sigma_t^{-1}\boldsymbol{\varepsilon}_t \tag{6.3}$$

The procedure to obtain the estimation involves Kalman Filter and Newton Raphson:

(1) Select initial values for the parameters $\Theta^{(0)} = \left\{\boldsymbol{\mu}_0^{(0)}, \Sigma_0^{(0)}, \Phi^{(0)}, Q^{(0)}, R^{(0)}\right\}$

(2) Run the Kalman Filter using $\Theta^{(0)}$ to get the innovations $\boldsymbol{\varepsilon}_t^{(0)}$ and covariance matrices $\Sigma_t^{(0)}$

(3) Run Newton Raphson algorithm to get updated parameters $\Theta^{(1)}$

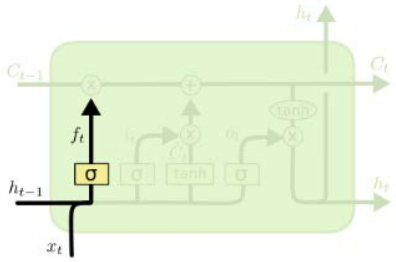
(4) Repeat (2) and (3) until the estimates stabilize.

6.3.2 Deep Learning Techniques

Recurrent Neural Network

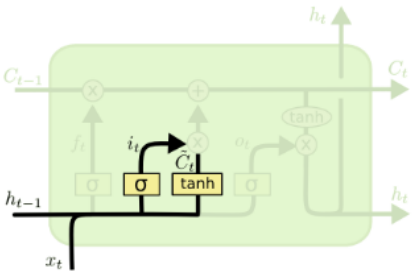
The Recurrent Neural Network (RNN) is a type of neural network developed for dealing with sequence data. In the trajectory context, the consecutive vehicle trajectories can be treated as sequence data in both spatial and temporal terms as indicated in the Kalman Filter model. An advantage that the RNN has is its flexibility. While only one step ahead prediction is allowed in

the KF prediction, the prediction step in RNN can be adjusted according to needs. The prediction term ranges from 1 step to several steps further. The RNN's another advantage over the tradition dense layer neural network is that during the training process it shares features learned across different positions of the sequence. Meanwhile comparing to dense layer neural network, the number of parameters are also greatly reduced. Many RNN models have been successfully applied in sequence data in the domain of speech and language modeling (Graves et al., 2013; Mikolov et al., 2010; Mikolov et al., 2011). Here one specific RNN is applied to the trajectory prediction, The Long Short Term Memory (LSTM). It is a member of the RNN family.



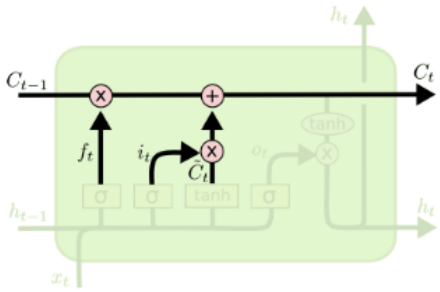
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

(a)



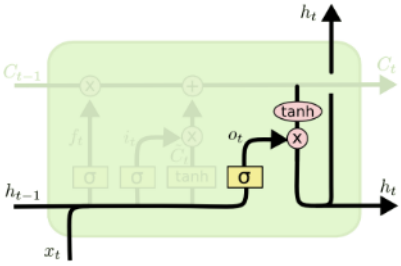
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

(b)



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

(c)



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

(d)

Figure 6. 3 Basic LSTM Structure

A LSTM cell is composed by 4 gates, which are the Forget Gate, the Input Gate, the Cell Gate and the Output Gate. A Forget is described as:

$$f_t = \sigma\left(W_f\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + b_f\right) \tag{6.4}$$

In which $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid activation function, $\boldsymbol{x}_t$ is the input vector at time t and it includes the BSM information. $\boldsymbol{h}_{t-1}$ is the hidden state vector passed from LSTM cell at time t-1. W_f and b_f together make a linear transformation of vector $[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]$. The function controls how much of last stage information are used at time t. Figure 6.3(a) (Olah, 2015) illustrate the forget gate's input and output direction.

An Input Gate is composed by two parts. The first part controls how much of current input information is used at time t, the second part calculates current stage's information. They can be described by:

$$i_t = \sigma\left(W_i\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + b_i\right) \tag{6.5}$$

$$\tilde{C}_t = \tanh\left(W_C\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + b_C\right) \tag{6.6}$$

In which W_i and b_i are linear transformation coefficients and i_t is the first part output limiting the amount of current information. Similarly, W_C and b_C are linear transformation coefficients for the second part and $\tilde{C}_t$ is interpreted as the current stage's information. The tangent function is expressed as $\tanh(x) = \left(e^{x} - e^{-x}\right)/\left(e^{x} + e^{-x}\right)$. This gate is illustrated in Figure 6.3(b).

A Cell Gate is a combination of previous Forget Gate and Input Gate, it describes how much information are kept from previous stage and current stage. It can be formulated as:

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \tag{6.7}$$

Where $\otimes$ represents the elementwise product of vectors. This gate is illustrated in Figure 6.3(c).

The last gate is the Output Gate, which calculates the hidden state vectors. The hidden state vectors may be passed down or used as output. This gate can be expressed as:

$$o_t = \sigma\left(W_o\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + b_o\right) \tag{6.8}$$

$$h_t = o_t \otimes \tanh\left(C_t\right) \tag{6.9}$$

In which W_o and b_o are linear transformation coefficients. This is described in Figure 6.3(d). Notice that f_t, i_t, $\tilde{C}_t$, C_t, o_t and h_t are of vector forms in implementation, which essentially increases the LSTM cell's dimension.

Bidirectional Structure

As introduced in the data overview section, besides the BSM data at current time t, the SPAT data is also available through the DSRC communication. In the specific intersection configuration case, the length of a vehicle extension call is 2 seconds. Therefore, for each phase, the not only past information is available, but the future information is also available. Based on previous LSTM

structure, a Bidirectional LSTM can actually create such a mechanism that the information can flow both forward and backward. The idea is originated from the paper by Mike Schuster and Kuldip K. Paliwal (Schuster and Paliwal, 1997). The illustration is shown in Figure 6.4. The bidirectional form comes from the fact two parallel LSTM cells are present at the same time. One is the norm LSTM cell which takes care of information flow following the sequence $1, 2, \cdots, n$ where n is the sequence total step number, indicated by the solid lines. The other one is a LSTM cell which processes the same sequence in a reversed order as indicated by the dashed lines. The outputs from both cells will be then concatenated for further operations, such as sum, product, etc.

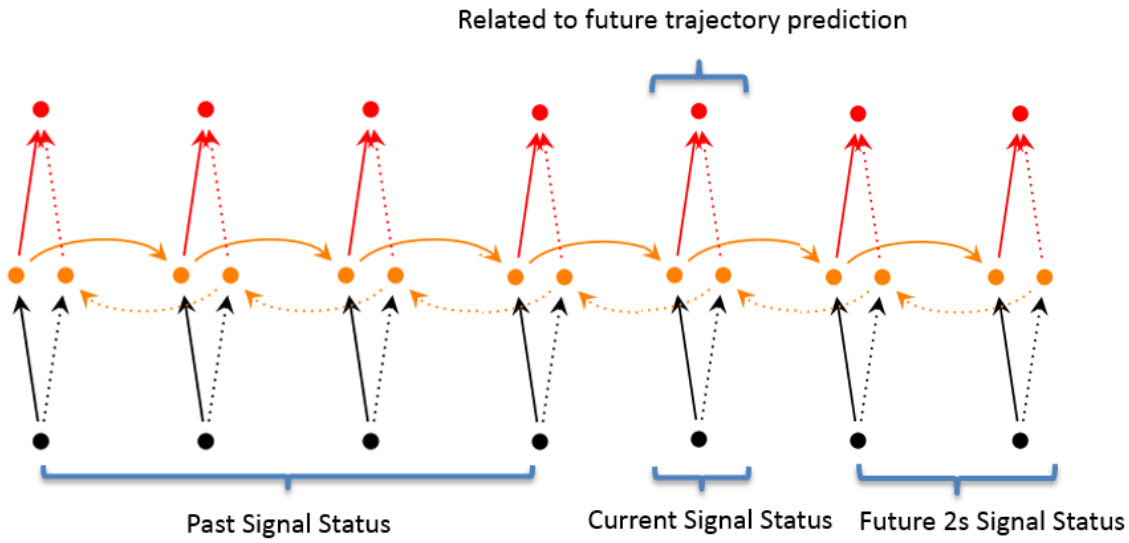


Figure 6. 4 Bidirectional Structure

Entity Embedding

The SPAT data belongs to the categorical type. Two normal ways for dealing with the categorical data are One Hot Encoding and Ordinal Encoding. The One Hot Encoding encodes the variable by

using Identity function expression, which returns a vector for each categorical variable. The Ordinal Encoding encodes the variable by using positive integers. These two techniques are popular in most machine learning problems. In the deep learning context, a superior way of encoding categorical variables is the Entity Embedding as proposed by Cheng Guo and Felix Berkhahn (Guo and Berkhahn, 2016). The principal of this encoding mechanism is to map each state of a categorical variable to a vector. Different from the One Hot Encoding's vector, this vector is trainable during the model training process, thus more flexible and accurate for specific application environments.

6.3.3 Feature Engineering

Besides the BSM and the SPAT information, other vehicle's information is also valuable for single vehicle's trajectory prediction. To find out surrounding related vehicles, a unsupervised learning technique, K-Means clustering method is used. The basic idea is to cluster vehicle trajectories into several group. Within each group, aggregated information can be extracted and further used as inputs to the neural network. Trajectory clustering is based on vehicle's spatial and temporal coordinates. Assume the spatial and temporal coordinate of a trajectory point is $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$ where x_1 is the local horizontal coordinate, x_2 is the local vertical coordinate and x_3 is the time stamp. The procedure of the clustering is as follows:

(1) Select k centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$ from all the trajectory points

(2) For each trajectory point $\boldsymbol{x}_i$, select g_i as this point's centroid by:

$$g_i = \arg\min_j \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_j \right\|_2^2 \tag{6.10}$$

(3) For each $\boldsymbol{\mu}_j$, update is made through:

$$\boldsymbol{\mu}_j = \frac{\sum_{g_i = j} \boldsymbol{x}_i}{\left| \{ i : g_i = j \} \right|} \tag{6.11}$$

(5) Repeat (2) and (3) until convergence and every trajectory point is labeled. Each vehicle's trajectory label is determined by the majority label of its points.

The iteration number for above procedure is determined by the within group variance. The value is usually selected as such for more iteration times after this value, the within group variances do not decrease significantly.

Using above clustering information, a single vehicle's environment features can be derived, including other vehicles dynamic information and their relations. A detailed feature table is presented in Table 6.1.

Table 6. 1 Feature Table

BSM Feature	SPAT Feature	Environment Feature
Local Horizontal Coordinate	Phase 1 Status	Average Horizontal Coordinate
Local Vertical Coordinate	Phase 2 Status	Average Vertical Coordinate
Speed	Phase 3 Status	Average Speed
Accelaration	Phase 4 Status	Average Acceleration
Requested Phase	Phase 5 Status	Minimal Horizontal Coordinate
Current Phase	Phase 6 Status	Minimal Vertical Coordinate
	Phase 7 Status	Minimal Speed
	Phase 8 Status	Minimal Acceleration
		Maximal Horizontal Coordinate
		Maximal Vertical Coordinate
		Maximal Speed
		Maximal Acceleration
		Vehicle Number
		Distance to Centroid Coordinate
		Difference between Current Local Horizontal Coordinate and Minimal Horizontal Coordinate
		Difference between Current Local Vertical Coordinate and Minimal Vertical Coordinate
		Difference between Current Local Horizontal Coordinate and Maximal Horizontal Coordinate
		Difference between Current Local Vertical Coordinate and Maximal Vertical Coordinate
		Difference between Speed and Minimal Speed
		Difference between Speed and Maximal Speed
		Difference between Acceleration and Minimal Acceleration
		Difference between Acceleration and Maximal Acceleration

6.3.4 Overall Structure

Two neural network structures are implemented in this research. One is a naïve LSTM neural work, which only uses BSM data for prediction. This is shown in Figure 6.5. This is an extended version of LSTM neural network. There are three layers of LSTM cell between input and output. For each LSTM cell, the inner hidden states are high dimensional (200 for each cell), as pointed in Section 6.3.2. Each second's BSM information $\boldsymbol{x}_i$ is used sequentially. Dropout modules are inserted among LSTM layers for the purpose of reducing overfitting. The mechanism of Dropout is to randomly mute certain proportions of neurons during training, setting their corresponding linear transformation coefficients to zero. Through multiple epochs of training of Stochastic Gradient Decent, the weights of each neuron's contribution to the output are expected to be distributed in a balanced form. Proportion of muted neurons is 0.4 in this case. After processing

5 seconds history trajectory data, the output will be the next 5 seconds vehicle position which is indicated by y_i. The three-layer structure, hyperparameters LSTM size and Dropout proportions are tuned carefully to achieve an optimal configuration.

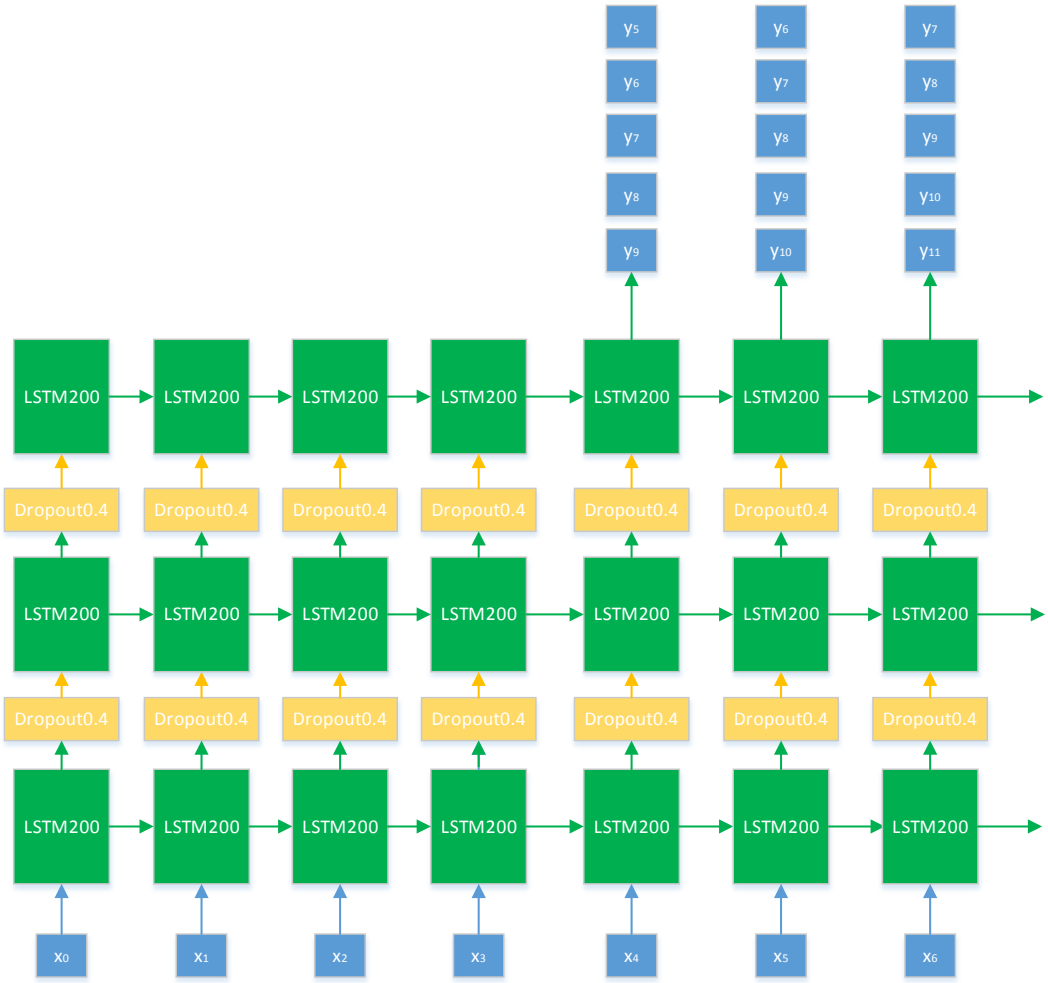


Figure 6. 5 Naïve LSTM Model

The other one is a more complex structure, composed of multiple inputs, utilizing all the features listed in Table 6.1. This is shown in Figure 6.6. The input is divided into three parts, the BSM information input, the SPAT information input and the Environment information input. For the

BSM part, three layer LSTM and one layer of dense layer are used. For the SPAT part, since there are 8 phases in the particular intersection, it is further divided into 8 parts. For each of the 8 inputs, the SPAT information is a 7 seconds sequence of categorical data. First the sequence is encoded by an Entity Embedding layer, followed by a batch normalization layer, a bidirectional LSTM layer and a flatten layer. A batch normalization layer works for reducing overfitting, the same purpose with Dropout's. The basic mechanism is that the layer takes previous layer's output (Entity Embedding layer in this case) and normalizes to a number with mean β and variance γ, which are then trainable parameters. This process is repeated for every batch of data during training process. The bidirectional LSTM was introduced in previous section. One thing to mention is that the following flatten layer is to concatenate the LSTM's output into a new vector. The Environment information is used an input and have same processing procedure with the BSM input. Then, three processed inputs are concatenated together. Finally, two dense layers with non-linear activation function "selu" (Klambauer et al., 2017) are used to add non-linearity for prediction. The "selu" function is in the form:

$$selu(x) = \lambda \begin{cases} x & x > 0 \\ \alpha e^{x} - \alpha & x \le 0 \end{cases} \tag{6.12}$$

The output layer is a dense layer with 10 neuron, representing future 5 second vehicle's positions in both local horizontal coordinates and vertical coordinates.

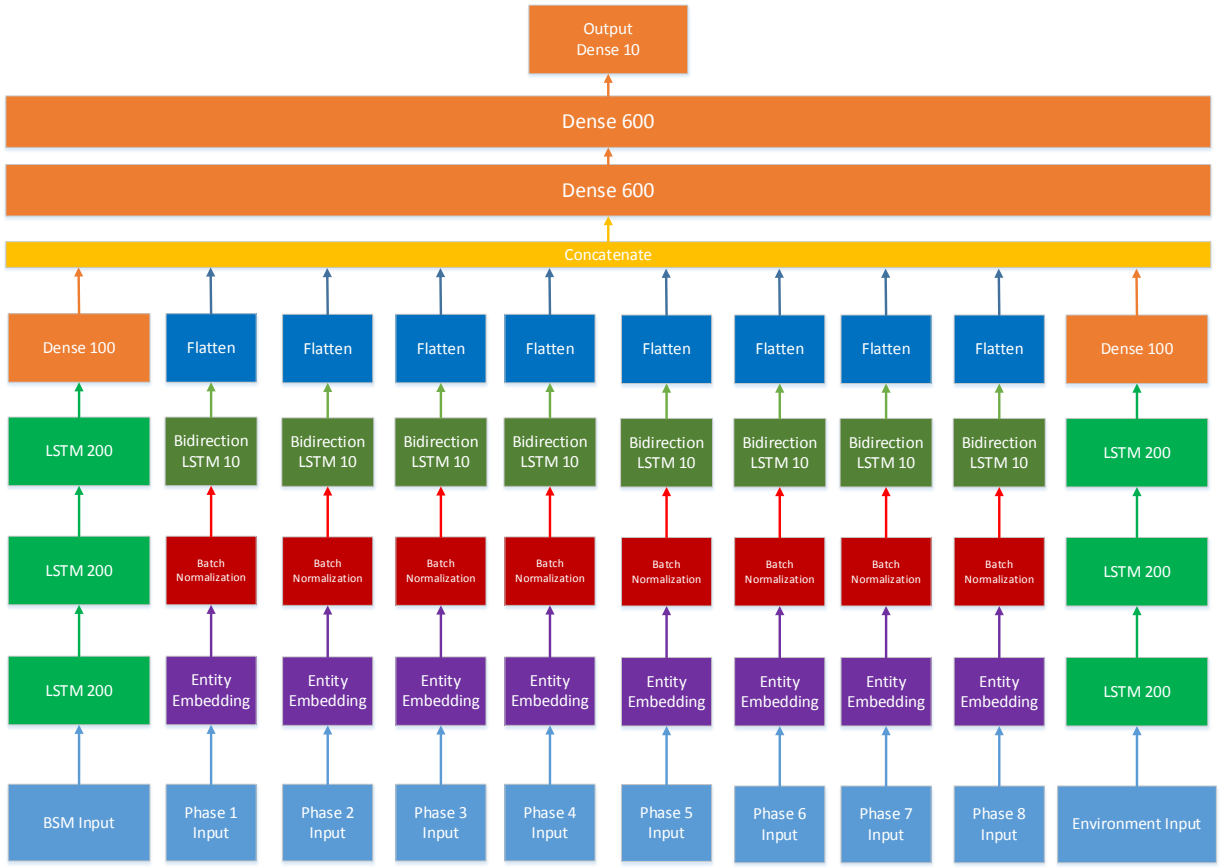


Figure 6. 6 Advanced LSTM Structure

The reason of implementing two structures is that by comparing prediction performances, the importance of incorporating SPAT and environment information will be highlighted. Both two neural networks are trained using Adam Mini Batch Stochastic Gradient Decent algorithm, using an Nvidia Titan Xp GPU. The hyperparameters for the training algorithm used in this research is listed in Table 6.2.

Table 6. 2 Training Hyperparameters

Parameter	Value
Optimization Algorithm	Adam
Loss Function	Mean Square Error
Learning Rate	0.001
Beta_1	0.9
Beta_2	0.999
Epsilon	1.00E-08
Decay	0

6.4 Results

6.4.1 Experiment Design

To evaluate the model performance, several factors are taken into considerations. The factors are the prediction step, the traffic volume, the vehicle type, the driver behavior and the vehicle speed. Among these 5 factors, only the prediction step and the vehicle type are fixed factors and the traffic volume, the driver behavior and the vehicle speed are treated as random factors. The step factor has 5 categories and they are Step 1 (the next first second), Step 2 (the next second second), Step 3 (the next third second), Step 4 (the next fourth second) and Step 5 (the next fifth second). The traffic volume has three categories, the low volume, the medium volume and the high volume. Three traffic volume levels are selected randomly: 200 veh/h on the main street and 100 veh/h on side streets for the low volume, 400 veh/h on the main street and 200 veh/h on side streets for the medium volume, 800 veh/h on the main street and 400 veh/h on side street for the high volume. The driver behavior has 3 categories and they are the conservative behavior, the normal behavior and the aggressive behavior (Habtemichael and Picado-Santos, 2013). Parameters describing the driver behaviors are listed in Table 6.3. The speed factor is in a continuous form. For every

combination of traffic volume, vehicle type and driver behavior, an one hour simulation was made in VISSIM to collect data. There are totally 442459 trajectories collected for the training purpose and another 5000 trajectories are used for test purpose. The specific intersection is selected from the MMITTS CV test corridor at Anthem, Arizona (Figure 6.7).

Table 6. 3 Parameter Settings for Driver Behavior

Driver Behavior Model	Model Parameters	Aggressive	Noraml	Conservative
Car Following Model (Wiedemann 99)	Standstill Distance	0.5 (m)	1.5 (m)	2.5 (m)
	Headway Time	4 (m)	4 (m)	4 (m)
	Following Variation	2 (m)	4 (m)	6 (m)
	Threshhold for Entering Following	-4 (m)	-8 (m)	-12 (m)
Lane Change Model (Free Lane Selection)	Maximum Decelartion (Trailing Vehicle)	-4 (m/s^2)	-4 (m/s^2)	-4 (m/s^2)
	Maximum Decelartion (Own)	-5 (m/s^2)	-3 (m/s^2)	-1 (m/s^2)
	Safety Distance Reduction Factor	0.1	0.6	1

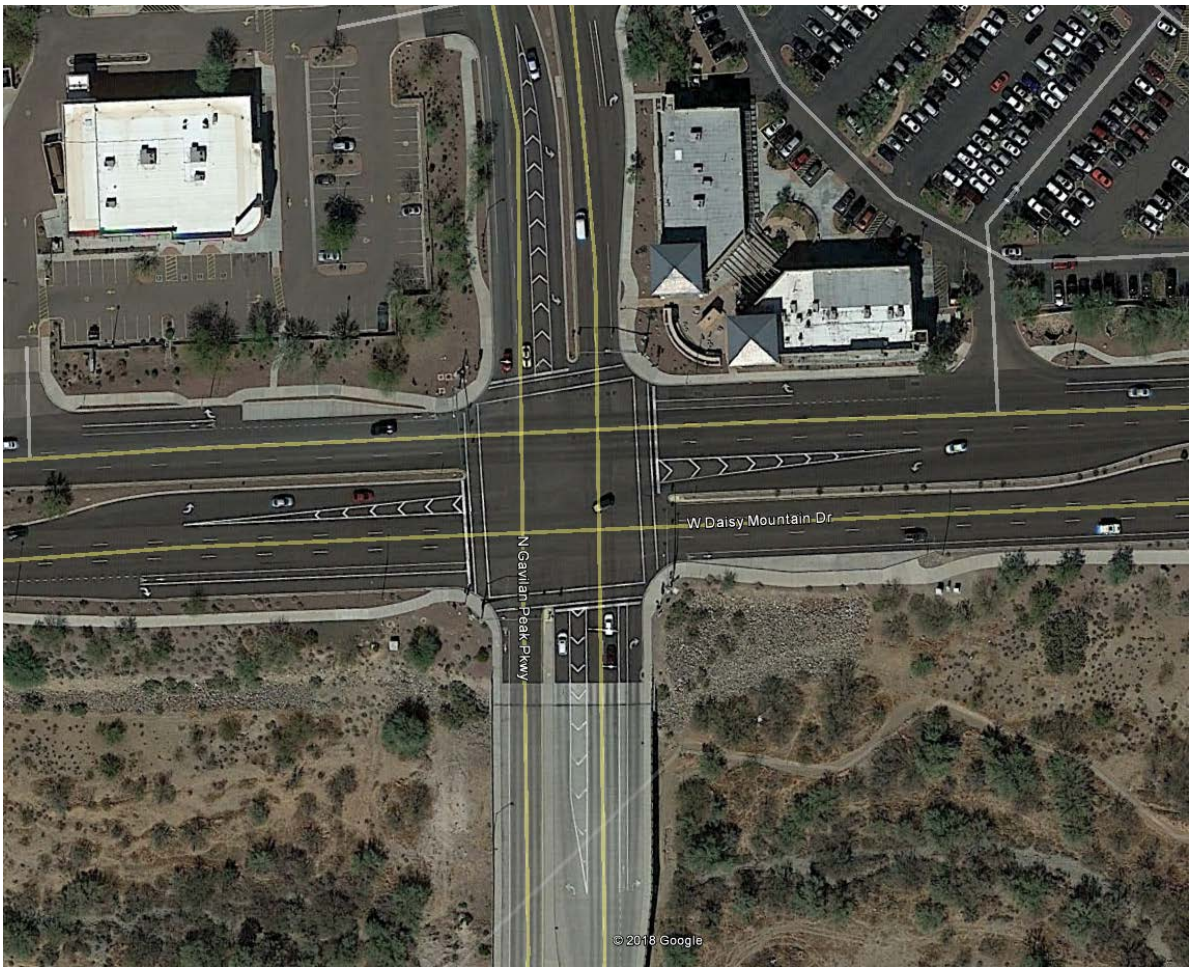


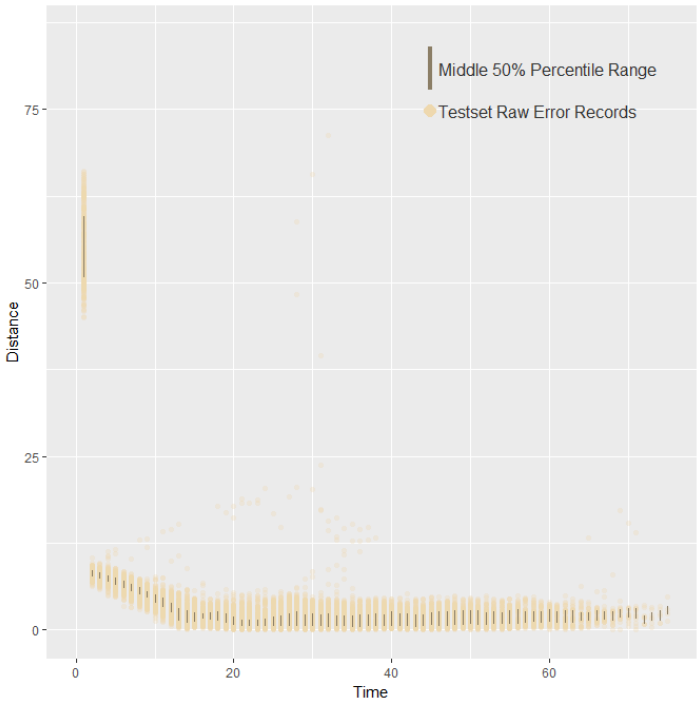
Figure 6. 7 Simulation Test Bed

6.4.2 Error Analysis and Model Comparison

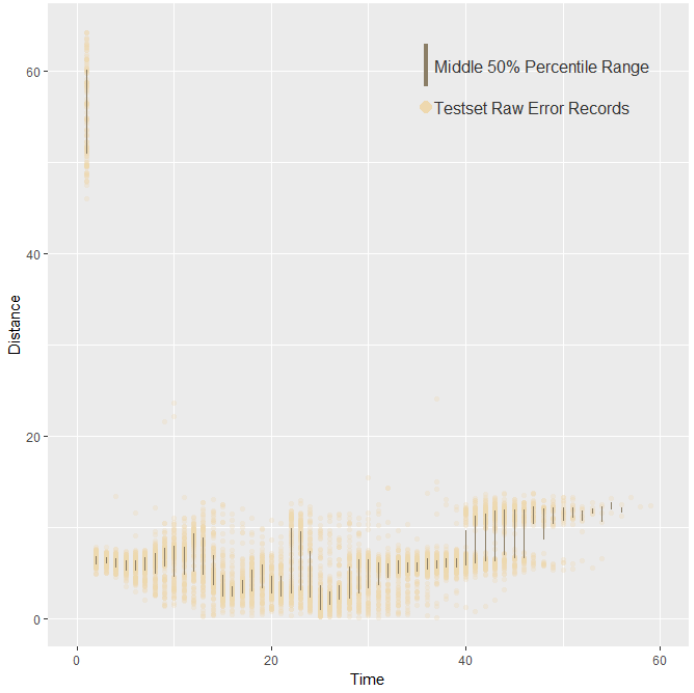
Intuitive Analysis

Error analysis by visual check is intuitive to understand how the predictions diverge from the real trajectories and will provide direction for further formal test. If the models differ greatly in this step, there is no need for further test. On the contrary, if the model difference is not easy to detect by eyes, statistical tests should give a better picture.

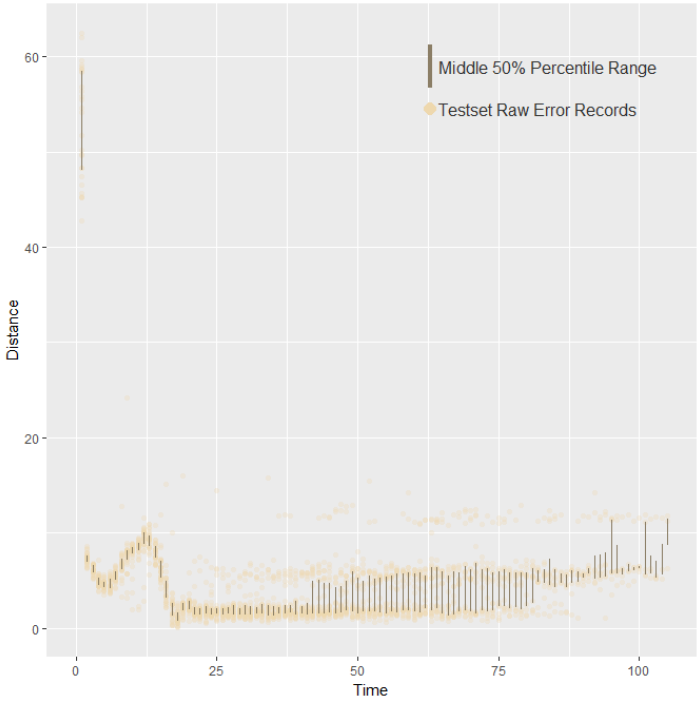
For the KF one step ahead prediction models, to achieve the best prediction accuracy, three sets of prediction models are trained respectively for the through vehicles, left turn vehicles and right turn vehicles. In Figure 6.8, for each case, the prediction errors in meter are plotted against the prediction time. It can be observed that for the through vehicle trajectories, the median prediction error is around 5 meters (Figure 6.8(a)). For the right turn and left turn vehicle trajectories, most of median prediction errors are between 5 meters and 10 meters (Figure 6.8(b) and Figure 6.8(c)).



(a)


Middle 50% Percentile Range
Testset Raw Error Records
Distance
Time
0
20
40
60

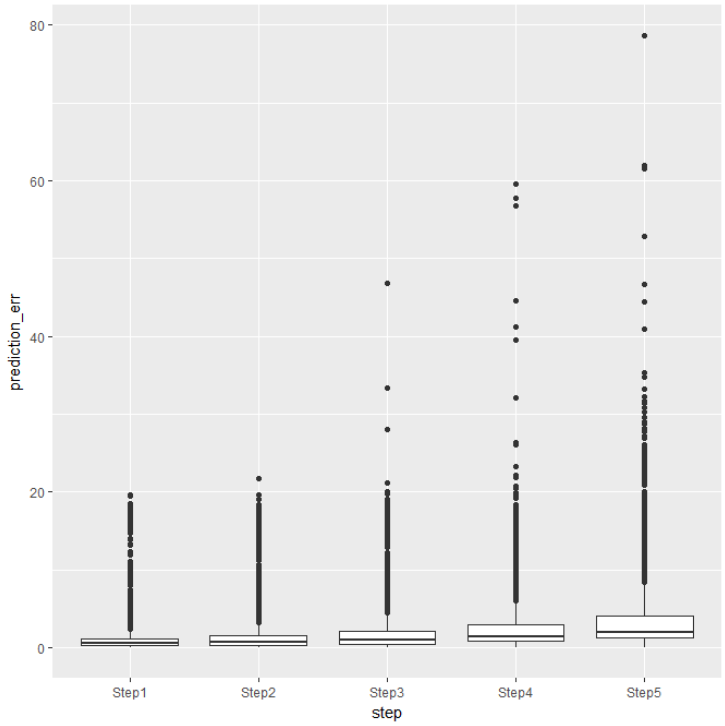
(b)


Middle 50% Percentile Range
Testset Raw Error Records
Distance
Time
0
20
40
60
0
25
50
75
100

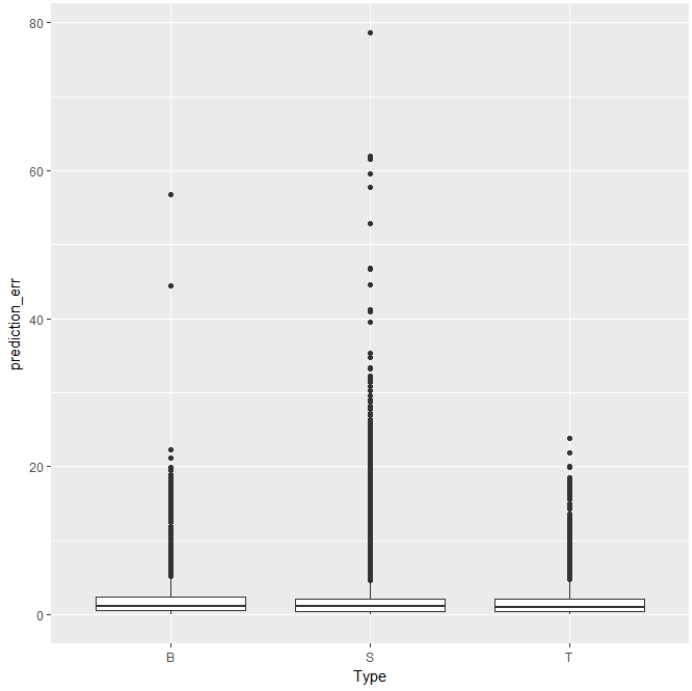
(c)

Figure 6. 8 Kalman Filter Prediction Error

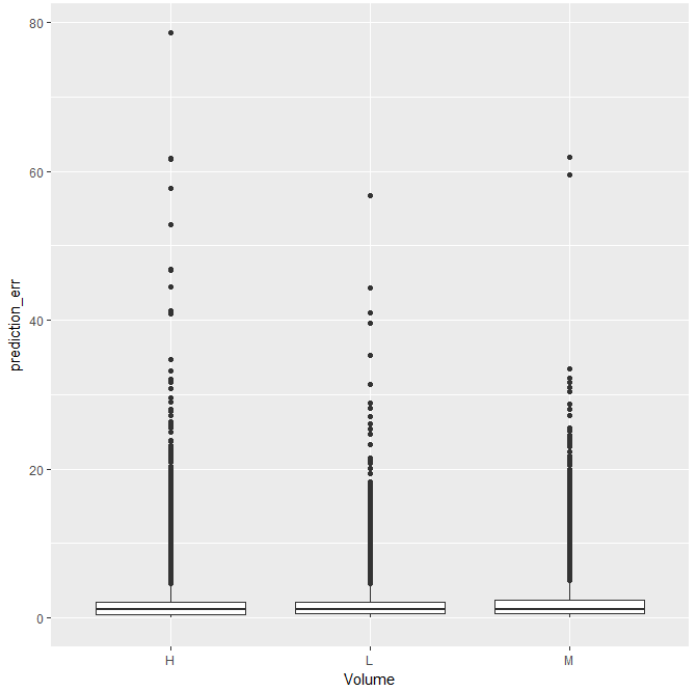
While the KF model prediction errors are seldom below 5 meters, the LSTM models perform much better (most median errors are around 1 meters). Figure 6.9 shows the prediction error for a naïve LSTM neural network by different factor groups. Figure 6.9(a) is a boxplot of prediction errors by steps. It seems that the prediction error is larger for further predictions. Figure 6.9(b) is a boxplot of prediction errors by vehicle type. The median errors for three vehicle types are almost same, but the sedan type obviously has more outliers and a larger error variance. Figure 6.9(c) is a boxplot of prediction errors by volume. The high volume group seems to have a larger error variance. Figure 6.9(d) is a boxplot of prediction errors by driver behavior and the conservative behavior group seems to have the smallest prediction error variance. Figure 6.9(e) is a 2 dimensional histogram with contours, describing the prediction error distribution by prediction error and speed. It can be observed that trajectories at low speed and high speed are of the largest proportion. Meanwhile it seems that prediction errors at high speed are slightly larger.


80
60
40
20
0
prediction_err
Step1
Step2
Step3
Step4
Step5
step

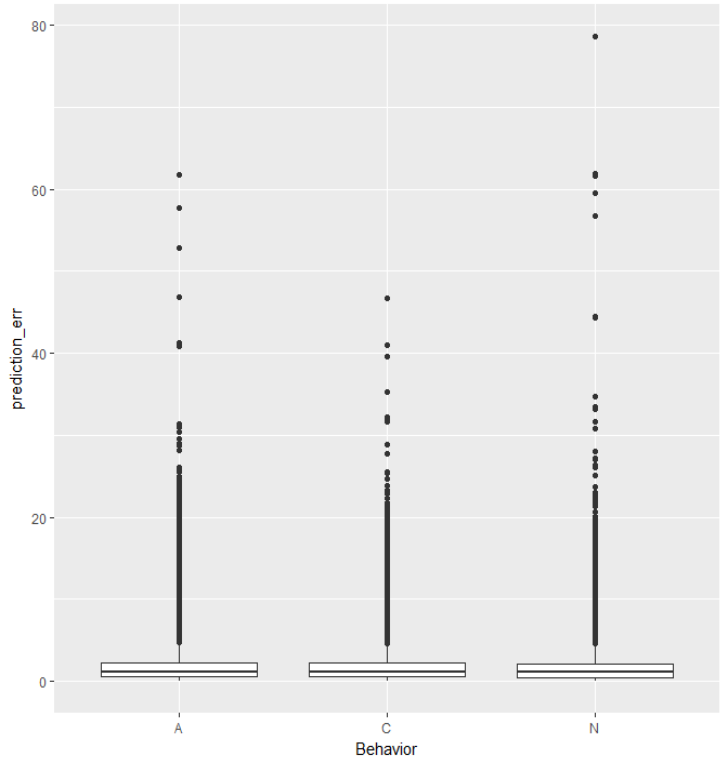
(a)


80
60
40
20
0
prediction_err
B
S
T
Type

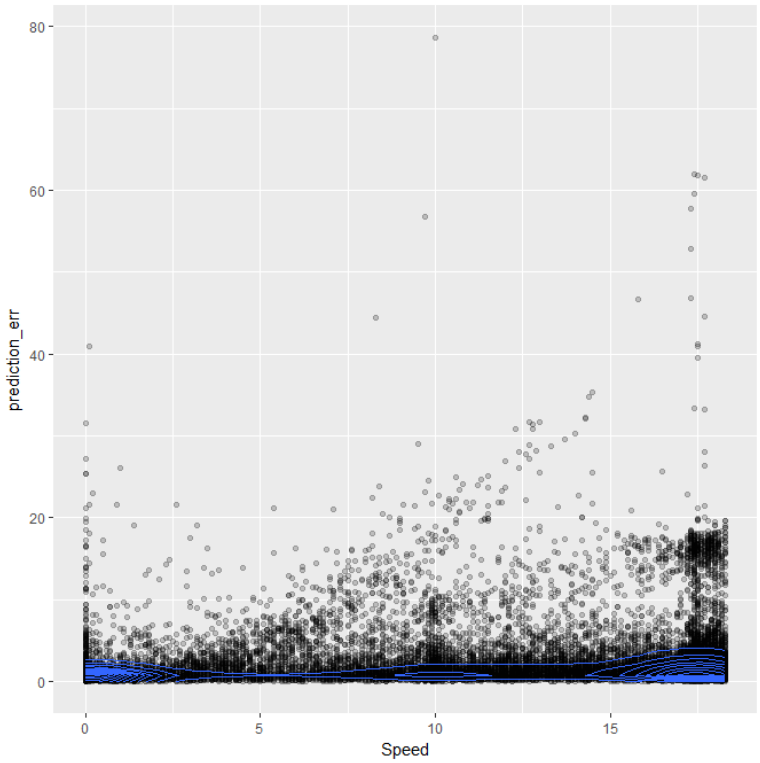
(b)


80
60
40
20
0
prediction_err
H
L
M
Volume

(c)


80
60
40
20
0
prediction_err
A
C
N
Behavior

(d)



(e)

Figure 6. 9 LSTM Prediction Errors

Restricted Maximum Likelihood Estimate

A statistical test procedure is utilized in this section for comparing two LSTM models, since it is more accurate to distinguish them than only by plots. By estimating the effects of both fixed factors and random factors, one can distinguish factors through their contributions to the prediction error. A commonly implemented method is the Analysis of Variance (ANOVA) procedure. But ANOVA estimates of variance components require that sample sizes are approximately balanced, with the number of observations for each set of conditions being almost equal. This cannot be satisfied in this research, as is indicated previously that the trajectory speed is not evenly distributed. An

alternative way is to use Maximum Likelihood Estimate (MLE) and Restricted Maximum Likelihood Estimate (RMLE). The MLE has one disadvantage with variance component estimation of the random factors. It assumes all fixed effects are known without error, therefore MLE often yields biased estimates for random factors when both fixed and random factors exist. On the other hand, REML estimators only maximize the portion of the likelihood that does not depend on the fixed effects. Thus, here REML is preferred for error analysis and model performance comparison.

The mixed effect model for describing prediction error is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}_1 + \boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 + \boldsymbol{e} \tag{6.13}$$

Where $\boldsymbol{y} = [y_1, y_2, \cdots, y_n]^T$ representing the prediction error vector including all the test prediction errors. X is the design matrix representing whether a certain fix effect exists or not for every prediction case. $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_7]$ represents the fix factor effects. $\boldsymbol{u}_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{in}]$ represents *ith* random factor's effect and further $\boldsymbol{u}_i \sim MVN(\boldsymbol{0}, \sigma_i^2 I)$. $\boldsymbol{e} = [e_1, e_2, \cdots, e_n]$ is the random noise and $\boldsymbol{e} \sim MVN(\boldsymbol{0}, \sigma_E^2 I)$. Thus, $\boldsymbol{y}$ is also multivariate normal with mean $X\boldsymbol{\beta}$ and variance-covariance matrix $V = \sum_{i=1}^{3} \sigma_i^2 I + \sigma_E^2 I$. In estimating the random factor's variance component in REML, the trick of removing fix effects is to have such a matrix K that $KX = 0$. Then, the mixed effect model becomes:

$$\boldsymbol{y}^{*} = K\boldsymbol{u}_1 + K\boldsymbol{\mu}_2 + K\boldsymbol{\mu}_3 + K\boldsymbol{e} \tag{6.14}$$

The transformed variance-covariance matrix $V^{*} = KVK^{T}$. The log likelihood is:

$$L = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|V^{*}\right| - \frac{1}{2}\left(\boldsymbol{y}^{*}\right)^{T}\left(V^{*}\right)^{-1}\boldsymbol{y}^{*} \tag{6.15}$$

In which the last term can be proved to be equal to $\boldsymbol{y}^{T}P\boldsymbol{y}$ (Searle and Casella), in which P is a function of original variance-covariance matrix V:

$$P = V^{-1} - V^{-1}X\left(X^{T}V^{-1}X\right)^{-1}X^{T}V^{-1} \tag{6.16}$$

By using Newton Raphson iterative method, the estimate of σ_i^2 and σ_E^2 can be obtained. Finally, the fixed effects $\boldsymbol{\beta}$ can be estimated by substituting $\hat{\sigma}_i^2$ and $\hat{\sigma}_E^2$ into the variance-covariance matrix:

$$\hat{\boldsymbol{\beta}} = \left(X^{T}\hat{V}^{-1}X\right)^{-1}X^{T}\hat{V}^{-1}\boldsymbol{y} \tag{6.17}$$

Both LSTM models are evaluated using above REML procedure. The details of comparison is listed in Table 6.4. Notice that in the advanced model's prediction, the speed and the driver behavior's variance components are much smaller than that of the naïve model. Meanwhile, the effect variances among steps are reduced. This is in accordance with their prediction performance. Table 6.5 gives the median prediction errors and variances for both models. Even though the advanced model has a slightly larger median error for the first and the second step prediction, its prediction error variance is much smaller than the naïve model. Obviously, the advanced model has a better ability for generalization and has better overall prediction results on the unseen data.

Table 6. 4 Factor Effect

Factor Type	Factor	Factor Level	Naïve Model	Advanced Model
Fixed	Intercept	Step 1+Bus	-0.210	-0.099
	Step	Step 2	0.193	0.058
		Step 3	0.536	0.254
		Step 4	0.916	0.509
		Step 5	1.300	0.788
	Vehicle Type	Sedan	0.001	0.105
		Truck	-0.104	0.110
Random	Vehicle Speed	NA	0.262^2	0.167^2
	Traffic Volume	NA	0.074^2	0.089^2
	Driver Behavior	NA	0.027^2	0.002^2

Table 6. 5 Stepwise Performance Comparison

Step	Naïve Model		Advanced Model	
	Median	Standard Deviation	Median	Standard Deviation
Step 1	0.62	1.92	0.83	1.65
Step 2	0.74	2.43	0.82	1.91
Step 3	1.00	2.97	0.97	2.31
Step 4	1.41	3.87	1.32	2.88
Step 5	2.02	5.04	1.78	3.47

6.5 Conclusion

Benefits of proposed prediction methods come from three aspects: First, the classic object tracking and prediction model, Kalman Filter, is implemented as a baseline model. The prediction results of this model is the starting line for the follow-up LSTM models. It is the motivation that advanced methods are needed for better prediction accuracy. Second, both naïve and advanced neural network structures are explored in this chapter. The naïve model only utilizes connected vehicle's BSM data, while the advanced model takes advantages of extra information that is available in the CV environment, the SPAT data and the environment data. The ways of incorporating new features provide the LSTM model the ability of considering multiple types of information sources at the same time and improve the model overall performance. Third, a REML based mixed effect error analysis model is developed for model validation and comparison. The error analysis essential helps find the model performance differences in different scenarios. By examining the prediction errors in both visual and statistical ways, model improvement can be made accordingly. Actually, the advanced neural network structure in this chapter is proposed after evaluating the disadvantage of the naïve structure.

7 Summary, Contributions and Future Research

7.1 Research Summary

Transportation Safety is considered of high priority for various public transportation system. With different data sources from public agencies and advanced Intelligent Transportation Systems (ITS), it is beneficial to learn lessons from past traffic accidents and more importantly, to initialize efforts targeting reducing such accidents. Many of the current transportation safety management systems suffering from integrating large amounts of data and effectively abstracting important information from it, especially when the data has various types and of high dimensions. Meanwhile, the development of Connected Vehicle provides an environment in which vehicle data can be made available for both road users and safety management systems. Such challenges and opportunities allows new perspectives and approaches for transportation safety analytics.

The main objective of this dissertation was to provide new methods of utilizing the state of art statistical machine learning and deep learning techniques to address the most concerned transportation safety issues. Those issues are:

7.1.1 Crash Accident Severity Analysis

Aiming at the crash severity investigation, this chapter provides an improved logistic regression based classification methodology utilizing modern statistical variable selection methods, case control subsampling technique and the machine learning ensemble algorithm. The overall framework is provide in Figure 7.1. The advantages of this proposed method are that it considers large numbers of potential causal factors and also it adapts well the common imbalanced structures of crash injury data. The 6-year truck crash data of the Interstate 10 Highway Arizona segment is studied. Proposed models are validated by using test datasets for injury severity prediction and

their performances are then evaluated by comparison. By interpreting the proposed models using the relative risk concept, critical factors leading to severe truck crash outcomes are identified including vehicle, roadway, lighting, weather, human and spatial factors. In this proposed analytics framework, user's interested factors can be examined through the quantified relative risk analysis. Based on the understanding of critical safety factors, prevention strategies to improve the national corridor safety are discussed.

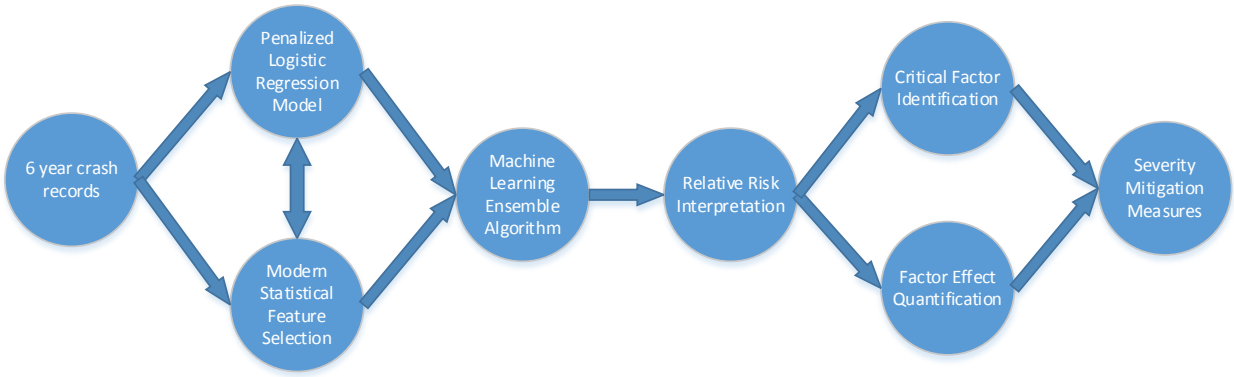


Figure 7. 1 Overall Framework for Accident Severity Analysis

7.1.2 Crash Accident Frequency Analysis

Travel time reliability and crash incident risk are the 2 major concerns for highway management policy makers and daily users. The relationship between traffic congestion and crash incidents is not widely investigated and existing studies have inconsistent conclusions. A corridor-level method is proposed to quantitatively assess the impacts of travel time reliability and weather factors on crash incident risk. The relationship among data and models are presented in Figure 2. High resolution data of travel time records and weather station observations are associated with time-of-day crash numbers of the Interstate 10 (I-10) Highway (Arizona part). Due to the high

missing rate of daily travel time records, a spatial temporal model is first employed for data imputation and the model is validated using an experiment of factorial design. The result shows this spatial temporal model outperforms the most used average imputation method in an environment where 20% original records are missing. Critical traffic segments and weather stations are identified using the Elastic Net (EN) statistical variable selection method based on the crash mean function which is derived in the Generalized Linear Model (GLM) scheme. The identification results are further validated using test datasets. The marginal effect analysis is conducted for the interpretation purpose. Finally, a series of mixed Poisson regression models with different heterogeneity forms are examined and the Chi-square Goodness of Fit test is performed to select the best fitted regression model. The conclusions show most identified critical segments' traffic congestion and increased wind speed are contributing to crash numbers, while few segments' traffic congestion is alleviating crash risk. Together with the mean function, the fitted mixed Poisson regression enables crash risk quantification and crash number prediction.

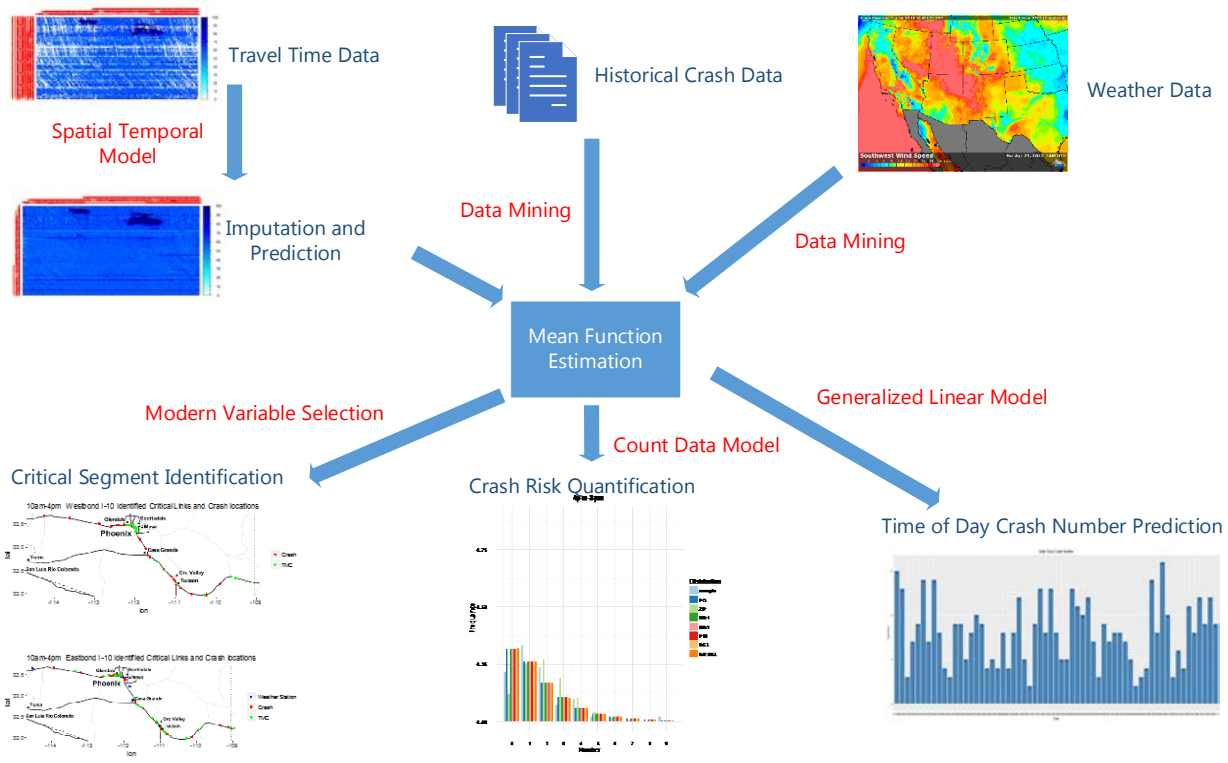


Figure 7. 2 Relationship among Data and Models in Accident Frequency Analysis

7.1.3 Vehicle Re-identification in a Connected Vehicle Environment

With the emergence of the connected vehicle technology estimating safety measures, such as conflict vehicles, utilizing vehicle trajectory data has received significant attention. High resolution location and motion data transferred from OBUs to RSU at intersection can be employed to re-identify a driver/vehicle and its particular route. The motivations for this topic are composed by 2 aspects: exploring the limits of current ID protection mechanism and providing ways of obtaining complete vehicle trajectories for research needs. This research explores the possibility of re-matching connected vehicles' ID using popular machine learning techniques, including Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Linear and Nonlinear Support Vector Machine (SVM) and Nearest Neighbor algorithms.

An experiment is conducted using a microscopic traffic simulation model through a Software-In-the-Loop technique. The best average mis-matching rate is 14%. To assess potential factors' effects on matching accuracy, a Poisson mixed regression model is analyzed under the Bayesian inference framework. Findings are: different matching algorithms vary in matching performance and the Linear SVM, the QDA and the LDA have the best accuracy results; traffic volume and market penetration rate have little impact on matching results; location and number of vehicles to be matched are considered significant. The results make the performance measure of future CV applications feasible and also suggest that more secure mechanisms are in need to protect the public.

7.1.4 Vehicle Trajectory Prediction in a Connected Vehicle Environment

Vehicle Trajectory prediction is made possible due to the Connected Vehicle technologies. Especially with the fast development of deep learning techniques in recent years, highly accurate prediction results could be achieved. By integrating connected vehicles' real time data, Basic Safety Message (BSM) data, Signal Phase and Timing (SPAT) data and surrounding vehicles dynamic data, this research develops effective methods for vehicle trajectory prediction and model validation. First, a Kalman Filter baseline model is implemented as a baseline model. It is the most commonly used method for moving object tracking and prediction. The best prediction error is around 5 meters. Then, two Recurrent Neural Netwok based models are come up. The first model is a naïve implementation of a Long Short Term Memory (LSTM) network. It only utilizes the vehicle's BSM data. Based on the prediction error analysis of the first model, an improved advanced LSTM model is developed to account for the intersection information and nearby vehicle dynamic information. Several deep learning modeling techniques are employed in developing the

second model, including the bidirectional structure, the entity embedding mechanism and the advanced feature engineering technique using unsupervised learning methods. The two LSTM models can outperform the classic Kalman Filter easily with the first step prediction error below 1 meter. A further detailed model validation framework is built on the factor effect analysis. Different potential factors are considered in the experiment design stage, including the traffic volume, vehicle type, driver behavior and vehicle speed. The Restricted Maximum Likelihood Estimate (RMLE) method is used to develop the mixed effect model and the result shows that the advanced LSTM model have better trajectory prediction performance than the naïve LSTM model. This research proves the potential capability of deep learning techniques being applied in the transportation safety analytics and even more complex autonomous driving scenarios.

7.2 Significance of Integrative Accident Analysis and Vehicle Trajectory Analysis

Accident analysis and vehicle trajectory analysis are two aspects of safety analysis. The former focuses on understanding the mechanisms of accident severity and frequency, and the latter focuses on real-time trajectory prediction. The integrative analysis of both aspects is beneficial, since the research efforts can contribute to each other.

The accident analysis reveals the critical safety factors in accident severity and risk assessment. Those identified factors and learned knowledges can be used for trajectory prediction research. For example, in the crash accident severity analysis, the accident characteristics of truck and non-truck are considered different according to the mosaic plot (Figure 3.1) and the Chi-square test. This figure highlights the over-represented and under-represented injury levels in each vehicle category. The result shows that the truck drivers and passengers' injuries have unique severity distribution from the non-truck ones, even though both of them have an extremely skewed

distribution. Such differences are probably due to the truck vehicle's unique properties and trajectory dynamics. Thus, including the vehicle type as a factor is a reasonable starting point for the vehicle trajectory analysis. Another critical factor identified is the speed. Using quantitative measures from the severity analysis, the high speed's impact on accident result is summarized in Table 3.4. According to this table, if the vehicle speed is between 70 mph and 80 mph , the relative risk of sever injury over no injury is increased by a factor of 2.92 and the relative risk of sever injury over possible injury is increased by a factor of 2.28. If the vehicle's speed is over 80 mph, it should be noticed that the relative risk of sever injury over no injury is increased by nearly 37 times. Besides, the speed over limit factor will also contribute to the relative risk increase slightly. It is clear that speeding and driving in high speed will increase the relative risk of being severely injured. Therefore, taking the speed factor into trajectory prediction model should be helpful in terms of giving drivers time to be aware of potential conflicts and crashes.

The developed trajectory prediction technique can be deployed at corridor's safety bottlenecks which are identified to likely have severer accidents and huge impacts on accident risks. The benefits are to reduce accident risk and alleviate potential accident severity. Figure 3.7 presents the identified critical I-10 segments which are likely contributing to sever accident injuries. The horizontal axis represents the longitude and the vertical axis represents the latitude. The solid black line is the shape of I-10 highway. Highlighted parts are segments identified risky. The risky degree is calculated according to the severity analysis model. The higher the degree, the more likely sever injuries happen. Figure 4.6 presents the identified I-10 segments whose travel time reliability is related to accident risks. The red dots represented recorded truck crashes and the green dots are called Traffic Management Channel (One can think it as representing highway segments). As can be seen in the figure, most crashes happened around the identified segments: the segment travel

time reliability has high impacts on crash risks. More specifically, for most segments, big travel time variances lead to high crash risks. The increased variance is usually presented as traffic jams in practice. In such situations, vehicle trajectories are likely to be affected by surrounding vehicle's dynamics. By using trajectory prediction in such segments is expected to reduce crash risks.

In sum, the integrative research of the two aspects is significant in terms of methodology approaches and developing safety improvement strategies. The accident analysis provides potentially important features to be considered in trajectory dynamic analysis, while the developed trajectory prediction technique has the potential to reduce crash accident severity and risks.

7.3 Research Contributions

This dissertation made several contributions to the current transportation safety analytics and provide the state of art statistical machine learning and deep learning technique implementations for transportation safety research. The main contributions include:

1. This research was the first in transportation literature to come up with methods for analyzing accident severity with consideration of the unique characteristic of history accident data. The proposed method can be applied to other similar accident datasets which are high dimensional and highly imbalanced. The contributed method improved the traditional machine learning techniques in terms of prediction accuracy and ease of interpretation.
2. This research developed a probability risk assessment framework for accident frequency analysis. The proposed framework has the capability of integrating multiple sources of data sources, taking account of the mutual impacts among corridor segments' travel time and

weather information. Meanwhile, a state space based spatial temporal model was come up to deal with the high proportions of missing records in travel time datasets.

3. This research was the first in transportation literature to explore the possibility of utilizing machine learning algorithms for re-identification connected vehicles in different intersections using partially observed trajectory information. The best proposed methods can achieve a high accuracy matching results in certain scenarios. A Bayesian based mixed effect model was employed in model validation and comparison, providing a new way for the traditional experiment design methodologies.
4. This research was the first in transportation literature to utilize the state of art advanced deep learning techniques and multiple transportation data sources for connect vehicle's trajectory prediction. The proposed method outperformed the classic "rule of thumb" object tracking and prediction method and the naïve deep learning approach.

In traditional transportation safety analytics, machine learning techniques are not widely applied and regarded as solutions to existing safety concerns. The proposed machine learning based approaches in this dissertation belong to such an effort that tries to solve the traditional transportation safety concerns from a novel point of view. The exploration of using statistical machine learning methods is an attempt to take advantages of the cutting edge statistical and computational techniques and apply them in transportation domains. Even though some progresses are obtained, there are still more challenges that need to be addressed. The success of proposed methods in solving generalized transportation issues depends on specific problems and how those problems are approached using proposed methods.

7.4 Future Research

Improvements to the proposed transportation safety analytics can be made in certain areas. Several suggestions may include:

1. *Exploring more variable selection methods for severity analysis*. Even though the improved classification model in this paper made some progress in prediction accuracy and interpretation ability, some more advanced variable selection methods could be utilized for identifying critical safety factors. For example, the weight mechanism of the Adaptive Elastic Net framework could be further explored. This could possibly increase the computation burden since it involves the topics of initial value and tuning parameter choice. Some pilot researches has made some interesting attempts (Algamal and Lee, 2015; Ghosh, 2011; Li et al., 2013).
2. *Collecting more data for crash risk assessment.* Due to the difficulties in obtaining quantitative data of human and vehicle factors, especially for those not involved in crash incidents, the current analysis approach does not incorporate above information either. This also partially explains the prediction error in Table 4.5 and Table 4.10 and the phenomenon that some crash incidents are far away from identified segments and weather stations in Figure 4.6. Future work may include data collection of such not considered factors and more sophisticated models for quantitative assessment purpose. This proposed crash frequency analysis model is promising in this sense, because there is no limit on the number of included explanatory variables, even greater than the data number. It renders the analysts to incorporate as many factors as they want.
3. *Exploring more complex vehicle re-identification scenarios.* One limitation when using simulation data is assuming that there are no new vehicles appearing in the test data set.

For instance, when a new vehicle enters the traffic stream from a parking lot between intersections, the mis-matching rate will likely increase. One mitigation measure to reduce the influence of newly appeared vehicles is to use static vehicle properties to do filtering. For example, if only 10 sedan vehicles appear in the first RSU range, then any other type of vehicle appearing at the next RSU must be a new vehicle which does not need ID matching. But such a mechanism cannot ensure new vehicles are filtered out every time given that they may have the same static properties as existing vehicles. The ultimate solution must be considered in the classification method itself. The Zero Shot Learning (ZSL) techniques (Xian et al., 2017) are promising in this regard. Future work can include such considerations and other matching techniques.

4. *Exploring more cutting edge techniques for trajectory prediction*. Even though the current connected vehicle trajectory prediction accuracy can be as high as 1 meter, some more advanced techniques could also be attempted. For example, the reinforcement learning techniques are promising in terms of deriving optimal control decisions in complex dynamic environment.

Appendix A

Feature	subcategory	Feature	subcategory
Age	Age≤20	Estimated Speed	Speed≤30
	20<Age≤30		30<Speed≤40
	30<Age≤40		40<Speed≤50
	40<Age≤50		50<Speed≤60
	50<Age≤60		60<Speed≤70
	60<Age≤70		70<Speed≤80
	Age>70		Speed>80
Sex	Female	Speed Over Limit	No
	Male		Yes
Safety Device	Not Applicable	Damage Area	AREA_1
	None Used		AREA_2
	Lap Belt		AREA_3
	Shoulder and Lap Belt		AREA_4
	Child Restraint System		AREA_5
	Helmet Used		AREA_6
	Air Bag Deployed		AREA_7
	Air Bag Deployed/Shoulder-Lap Belt		AREA_8
			AREA_9
Violation	No Improper Action		NONE
	Speed Too Fast for Conditions		TOTALED
	Exceeded Lawful Speed		UNDERCARRIAGE
	Followed Too Closely		
	Disregarded Traffic Signal	Control Type	Flashing Traffic Control Signal
	Made Improper Turn		No Controls
	Drove Rode in Opposing Traffic Lane		Person Flagger Law Enforcement Xing Guard ETC
	nowingly Operated with Faulty Missing Equipmen		Railroad Crossing Device
	Passed in No Passing Zone		Stop Signs
	Unsafe Lane Change		Traffic Control Signal
	Failed to Keep in Proper Lane		Warning Signs
	Other Unsafe Passing		Yield Signs
	Inattention Distraction		
	Electronic Communications Device	Surface Condition	Dry
	Failed to Yield Right of Way		Ice Frost
			Oil
Road Alignment	Curve Left		Slush
	Curve Right		Snow
	Straight		Water Standing Moving
			Wet
Road Grade	Downhill	Incident Year	
	Hillcrest		2010
	Level		2011
	Sag Bottom		2012
	Uphill		2013
			2014
Unit Action	Avoiding Vehicle Object Pedestrain		2015
	Backing	Incident Month	
	Changing Lanes		1
	Driverless Moving Vehicle		2
	Going Straight Ahead		3
	Leaving Parking Position		4
	Making Left Turn		5
	Making Right Turn		6
	Negotiating a Curve		7
	Overtaking Passing		8
	Properly Parked		9
	Slowing in Traffiway		10
	Stopped in Trafficway		11
	Working on Road		12
Body Style	PASSENGER_12PU_PICKUP_1_2_TON	Incident Day of Week	
	PASSENGER_34PU_PICKUP_3_4_TON		1
	PASSENGER_PU_PICKUP		2
	TRUCK_1TPU_PICKUP_1_TON		3
	TRUCK_1TVN_VAN_1_TON		4
	TRUCK_AR_ARMORED_TRUCK		5
	TRUCK_BS_BUS		6
	TRUCK_CB_CAB_CHASSIS		7
	TRUCK_CM_CONCRETE_OR_TRANSIT_MIXER	Incident Hour	
	TRUCK_CR_CRANE		0
	TRUCK_DP_DUMP_TRUCK		1
	TRUCK_DRTK_DRILLING_TRUCK		2
	TRUCK_FB_FLATBED_OR_PLATFORM		3
	TRUCK_FT_FIRE_TRUCK		4
	TRUCK_GG_GARBAGE_OR_REFUSE		5
	TRUCK_GR_GLASS_RACK		6
	TRUCK_PN_PANEL		7
	TRUCK_RF_REFRIGERATED_VAN		8
	TRUCK_SCBS_SCHOOL_BUS		9
	TRUCK_SR_SERVICE_BODY_TRUCK		10
	TRUCK_ST_STAKE_OR_RACK		11
	TRUCK_TK_TRUCK		12
	TRUCK_TN_TANK		13
	TRUCK_TRTK_TRENCH_TRUCK		14
	TRUCK_TT_TRUCK_TRACTOR		15
	TRUCK_VN_VAN		16
	TRUCK_WR_TOW_TRUCK_WRECKER		17
Make			18
	CADILLAC		19
	CHEVROLET		20
	CHRYSLER		21
	DODGE		22
	EAGLE		23
	FORD	Collison Manner	
	FREIGHTLINER		Single Vehicle
	GM		Angle(front to side, other than left turn)
	GMC		Left Turn
	HINO		Rear End
	HONDA		Head On
	HYUNDAI		Sideswipe Same Direction
	INTERNATIONAL		Sideswip Opposite Direction
	ISUZU		Rear to Side
	JEEP		Rear to Rear
	KIA	Light Condition	
	KENWORTH		Daylight
	LINCOLN		Dawn
	MACK		Dusk
	MAZDA		Dark Lighted
	MERCURY		Dart Not Lighted
	MERCEDES_BENZ		Dark Unknown Lighting
	MITSUBISHI	County ID	
	NISSAN		Cochise
	OLDSMOBILE		La Paz
	PLYMOUTH		Maricopa
	PONTIAC		Pima
	PETERBILT		Pinal
	STERLING	Traffic Way Type	
	TOYOTA		One Way Trafficway
	TRIUMPH		Two Way Not Divided
	VOLKSWAGEN		Two Way Not Divided with Continous Left Turn Lane
	VOLVO		Two Way Divided Unprotected Painted 4 Feet Median
Color			Two Way Divided Positive Median Barrier
	Beige	Intersection Type	
	Black		Not an Intersection
	Blue		Four Way Intersection
	Brown		T Intersection
	Bronze		Y Intersection
	Chrome		Intersection as Part of Interchange
	Copper		Traffic Circle
	Cream	Weather	
	Blue Dark		Clear
	Green Dark		Cloudy
	Gold		Sleet Hail Freezing Rain or Drizzle
	Green		Rain
	Gray		Snow
	Lavender Purple		Severe Crosswinds
	Blue Light		Blowing Sand Soil Dirt
	Burgundy Purple		Fog Smog Smoke
	Maroon	Latitude	Blowing Snow
	Multicolored		
	Orange		31.963
	Purple		31.964
	Pink		31.965
	Red		...
	Aluminum		33.678
	Silver	Longitude	
	Tan		-114.531
	Teal Green		-114.515
	Turquoise Blue		-114.510
	White		...
	Yellow		-109.131

Appendix B

For a state space model specified by:

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \Upsilon \boldsymbol{\mu}_t + \boldsymbol{w}_t \tag{B.1}$$

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \Gamma \boldsymbol{\mu}_t + \boldsymbol{v}_t \tag{B.2}$$

Where $\boldsymbol{\mu}_t$ the vector of inputs at t, the other letter definitions is are same with that in formula 4.2 and 4.3.

Property 1 The Kalman Filter

With initial conditions $\boldsymbol{x}_0^0 = \boldsymbol{\mu}_0$ and $P_0^0 = \Sigma_0$, for $t = 1, 2, \cdots, n$

$$\boldsymbol{x}_t^{t-1} = \Phi \boldsymbol{x}_{t-1}^{t-1} + \Upsilon \boldsymbol{\mu}_t \tag{B.3}$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi^t + Q \tag{B.4}$$

With

$$\boldsymbol{x}_t^t = \boldsymbol{x}_t^{t-1} + K_t \left(\boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{\mu}_t \right) \tag{B.5}$$

$$P_t^t = \left[I - K_t A_t \right] P_t^{t-1} \tag{B.6}$$

Where

$$K_t = P_t^{t-1} A_t^T \left[A_t P_t^{t-1} A_t^T + R \right]^{-1} \tag{B.7}$$

Is called the Kalman gain. Prediction for $t > n$ is accomplished via B.3 and B.4 with initial conditions $\boldsymbol{x}_n^n$ and P_n^n.

Property 2 The Kalman Smoother

With initial conditions $\boldsymbol{x}_n^n$ and P_n^n obtained from Property 1, for $t = n, n-1, \cdots, 1$

$$\boldsymbol{x}_{t-1}^n = \boldsymbol{x}_{t-1}^{t-1} + J_{t-1} \left(\boldsymbol{x}_t^n - \boldsymbol{x}_t^{t-1} \right) \tag{B.8}$$

$$P_{t-1}^n = P_{t-1}^{t-1} + J_{t-1} \left(P_t^n - P_t^{t-1} \right) J_{t-1}^T \tag{B.9}$$

Where

$$J_{t-1} = P_{t-1}^{t-1} \Phi^T \left[P_t^{t-1} \right]^{-1} \tag{B.10}$$

References

ADOT, 2015. Arizona Crash Data Set, https://www.azdot.gov/.

ADOT, 2017. Arizona's Crash Report Forms Instruction Manual, 11th ed.

Aguero-Valverde, J., 2013. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. *Accident Analysis & Prevention* 50, 289-297.

Algamal, Z.Y., Lee, M.H., 2015. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in biology and medicine* 67, 136-145.

Anderson, J.A., 1972. Separate sample logistic discrimination. *Biometrika* 59(1), 19-35.

Antonucci, N.D., Hardy, K.K., Slack, K.L., Pfefer, R., Neuman, T.R., 2004. NCHRP Report 500: Guidance for Implementation of the AASHTO Strategic Highway Safety Plan. Volume 12: A Guide for Reducing Collisions at Signalized Intersections. *Transportation Research Board of the National Academies, Washington, DC*.

Armstrong, J.S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting* 8(1), 69-80.

Barceló, J., Montero, L., Marqués, L., Carmona, C., 2010. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*(2175), 19-27.

Barth, M., Farrell, J.A., 1999. The Global Positioning System & Inertial Navigation. *McGraw-Hill* 8, 21-56.

Behnood, A., Mannering, F.L., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: some empirical evidence. *Analytic methods in accident research* 8, 7-32.

Bellman, R., 1961. Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ*.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24(3), 127-135.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. *Technometrics* 37(4), 373-384.

Breiman, L., 1996. Bagging predictors. *Machine learning* 24(2), 123-140.

Brennan Jr, T.M., Ernst, J.M., Day, C.M., Bullock, D.M., Krogmeier, J.V., Martchouk, M., 2010. Influence of vertical sensor placement on data collection efficiency from bluetooth MAC address collection devices. *Journal of Transportation Engineering* 136(12), 1104-1109.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* 88(421), 9-25.

Brimley, B., Saito, M., Schultz, G., 2012. Calibration of Highway Safety Manual safety performance function: development of new models for rural two-lane two-way highways. *Transportation Research Record: Journal of the Transportation Research Board*(2279), 82-89.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* 7(4), 434-455.

Cameron, A.C., Trivedi, P.K., 2013. *Regression analysis of count data*. Cambridge university press.

Carter, P.M., Flannagan, C.A., Reed, M.P., Cunningham, R.M., Rupp, J.D., 2014. Comparing the effects of age, BMI and gender on severe injury (AIS 3+) in motor-vehicle crashes. *Accident Analysis & Prevention* 72, 146-160.

Chang, L.-Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science* 43(8), 541-557.

Chang, L.-Y., Mannering, F., 1999. Analysis of injury severity and vehicle occupancy in truck-and non-truck-involved accidents. *Accident Analysis & Prevention* 31(5), 579-592.

Charbonnier, S., Pitton, A.-C., Vassilev, A., 2012. Vehicle re-identification with a single magnetic sensor, *Instrumentation and Measurement Technology Conference (I2MTC), 2012 IEEE International*. IEEE, pp. 380-385.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single-and multi-vehicle accidents on rural highways. *Accident Analysis & Prevention* 43(5), 1677-1688.

Chira-Chavala, T., Yoo, S., 1994. Potential safety benefits of intelligent cruise control systems. *Accident Analysis & Prevention* 26(2), 135-146.

Corbeil, R.R., Searle, S.R., 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* 18(1), 31-38.

Cressie, N., Wikle, C.K., 2015. *Statistics for spatio-temporal data*. John Wiley & Sons.

Cryer, J.D., Chan, K.-s., 2008. Time series analysis with applications in R, *Springer texts in statistics*, 2nd ed. Springer, New York, pp. 1 online resource (xiii, 491 p.).

De Brébisson, A., Simon, É., Auvolat, A., Vincent, P., Bengio, Y., 2015. Artificial neural networks applied to taxi destination prediction. *arXiv preprint arXiv:1508.00021*.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014a. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis & Prevention* 70, 320-329.

Dong, C., Richards, S.H., Clarke, D.B., Zhou, X., Ma, Z., 2014b. Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression. *Safety science* 70, 63-69.

Endo, Y., Nishida, K., Toda, H., Sawada, H., 2017. Predicting Destinations from Partial Trajectories Using Recurrent Neural Network, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 160-172.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 96(456), 1348-1360.

FHWA, 2015. National Performance Management Research Data Set, https://ops.fhwa.dot.gov/freight/freight_analysis/perform_meas/vpds/npmrdsfaqs.htm.

FHWA, 2017. FHWA Announces Vehicle-to-Infrastructure Guidance, in: Administration, F.H. (Ed.). Federal Highway Administration.

Fithian, W., Hastie, T., 2014. Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics* 42(5), 1693.

Friedman, J., Hastie, T., Simon, N., Tibshirani, R., 2016. glmnet: Lasso and elastic-net regularized generalized linear models, R-package version 2.0–5.

Friedman, J., Hastie, T., Tibshirani, R., 2001. *The elements of statistical learning*. Springer series in statistics Springer, Berlin.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1), 1.

Friendly, M., 1994. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* 89(425), 190-200.

García-Cortés, L.A., Sorensen, D., 2001. Alternative implementations of Monte Carlo EM algorithms for likelihood inferences. *Genetics Selection Evolution* 33(4), 443.

Geisser, S., 1975. The predictive sample reuse method with applications. *Journal of the American statistical Association* 70(350), 320-328.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.

Ghosh, S., 2011. On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing* 21(3), 451-462.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice* 1, 19.

Graves, A., Mohamed, A.-r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks, *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, pp. 6645-6649.

Guo, C., Berkhahn, F., 2016. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

Habtemichael, F., Picado-Santos, L., 2013. Sensitivity analysis of VISSIM driver behavior parameters on safety of simulated vehicles and their interaction with operations of simulated traffic, *92nd Annual Meeting of the Transportation Research Board, Washington, DC*.

Hadfield, J.D., 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software* 33(2), 1-22.

Haghani, A., Hamedi, M., Sadabadi, K., Young, S., Tarnoff, P., 2010. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board*(2160), 60-68.

Hainen, A., Wasson, J., Hubbard, S., Remias, S., Farnsworth, G., Bullock, D., 2011. Estimating route choice and travel time reliability with field observations of Bluetooth probe vehicles. *Transportation Research Record: Journal of the Transportation Research Board*(2256), 43-50.

Hofleitner, A., Herring, R., Bayen, A., 2012. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological* 46(9), 1097-1122.

Holmes, E., Ward, E., Wills, K., 2013. MARSS: Multivariate autoregressive state-space modeling. *R package version* 3(9).

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47(260), 663-685.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International journal of forecasting* 22(4), 679-688.

Islam, S., Jones, S.L., Dye, D., 2014. Comprehensive analysis of single-and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. *Accident Analysis & Prevention* 67, 148-158.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation planning and Technology* 15(1), 41-58.

Jovanis, P.P., Chang, H.-L., 1986. Modeling the relationship of accidents to miles traveled. *Transportation Research Record* 1068, 42-51.

Juliana Gruenwald Henderson, K.J., 2017. Connected Cars: Privacy, Security Issues Related to Connected, Automated Vehicles.

Kavaler, R., Kwong, K., Raman, A., Varaiya, P., Xing, D., 2011. Arterial performance measurement system with wireless magnetic sensors, *ICTIS 2011: Multimodal Approach to Sustained Transportation System Development: Information, Technology, Implementation*, pp. 377-385.

Kenney, J.B., 2011. Dedicated short-range communications (DSRC) standards in the United States. *Proceedings of the IEEE* 99(7), 1162-1182.

Kim, J.-K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention* 50, 1073-1081.

Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S., 2017. Self-normalizing neural networks, *Advances in Neural Information Processing Systems*, pp. 971-980.

Kogut, G.T., Trivedi, M., 2001. Maintaining the identity of multiple vehicles as they travel through a video network, *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*. IEEE, pp. 756-761.

Kononov, J., Bailey, B., Allery, B., 2008. Relationships between safety and both congestion and number of lanes on urban freeways. *Transportation Research Record: Journal of the Transportation Research Board*(2083), 26-39.

Kwon, J., Barkley, T., Hranac, R., Petty, K., Compin, N., 2011. Decomposition of travel time reliability into various sources: incidents, weather, work zones, special events, and base capacity. *Transportation Research Record: Journal of the Transportation Research Board*(2229), 28-33.

Kwong, K., Kavaler, R., Rajagopal, R., Varaiya, P., 2009. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies* 17(6), 586-606.

LeCun, Y., Touresky, D., Hinton, G., Sejnowski, T., 1988. A theoretical framework for back-propagation, *Proceedings of the 1988 connectionist models summer school*. CMU, Pittsburgh, Pa: Morgan Kaufmann, pp. 21-28.

Li, J., Jia, Y., Zhao, Z., 2013. Partly adaptive elastic net and its application to microarray classification. *Neural Computing and Applications* 22(6), 1193-1200.

Li, P., Souleyrette, R.R., 2016. A Generic Approach to Estimate Freeway Traffic Time Using Vehicle ID-Matching Technologies. *Computer-Aided Civil and Infrastructure Engineering* 31(5), 351-365.

Liu, C., Subramanian, R., 2009. Factors related to fatal single-vehicle run-off-road crashes.

Lomax, T., Schrank, D., Turner, S., Margiotta, R., 2003. Selecting travel reliability measures. Texas Transportation Institute, Cambridge systematics. Inc.

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.

MacCarley, C.A., 2001. Video-based vehicle signature analysis and tracking system phase 2: algorithm development and preliminary testing. *California Partners for Advanced Transit and Highways (PATH)*.

Meng, Q., Qu, X., 2012. Estimation of rear-end vehicle crash frequencies in urban road tunnels. *Accident Analysis & Prevention* 48, 254-263.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention* 26(4), 471-482.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S., 2010. Recurrent neural network based language model, *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., Khudanpur, S., 2011. Extensions of recurrent neural network language model, *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, pp. 5528-5531.

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

Montgomery, D.C., 2005. *Introduction to statistical quality control*, 5th ed. John Wiley, Hoboken, N.J.

Montgomery, D.C., 2017. *Design and analysis of experiments*. John Wiley & Sons.

Morton, J., Wheeler, T.A., Kochenderfer, M.J., 2017. Analysis of recurrent neural networks for probabilistic modeling of driver behavior. *IEEE Transactions on Intelligent Transportation Systems* 18(5), 1289-1298.

Nelder, J.A., Baker, R.J., 1972a. Generalized linear models. *Encyclopedia of statistical sciences*.

Nelder, J.A., Baker, R.J., 1972b. *Generalized linear models*. Wiley Online Library.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. *Applied linear statistical models*. Irwin Chicago.

Noland, R.B., Quddus, M.A., 2005. Congestion and safety: A spatial analysis of London. *Transportation Research Part A: Policy and Practice* 39(7), 737-754.

Olah, C., 2015. Understanding LSTM Networks.

Pecher, P., Hunter, M., Fujimoto, R., 2016. Data-Driven Vehicle Trajectory Prediction, *Proceedings of the 2016 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. ACM, pp. 13-22.

Pinheiro, J.C., Chao, E.C., 2006. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics* 15(1), 58-81.

Press, S.J., Wilson, S., 1978. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association* 73(364), 699-705.

Prevost, C.G., Desbiens, A., Gagnon, E., 2007. Extended kalman filter for state estimation and trajectory prediction of a moving object detected by an unmanned aerial vehicle, *American Control Conference, 2007. ACC'07*. IEEE, pp. 1805-1810.

Pu, W., 2011. Analytic relationships between travel time reliability measures. *Transportation Research Record: Journal of the Transportation Research Board*(2254), 122-130.

Quayle, S., Koonce, P., DePencier, D., Bullock, D., 2010. Arterial performance measures with media access control readers: Portland, Oregon, pilot study. *Transportation Research Record: Journal of the Transportation Research Board*(2192), 185-193.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Richardson, J., Smith, B., Fontaine, M., Turner, S., 2011. Network stratification method by travel time variation. *Transportation Research Record: Journal of the Transportation Research Board*(2256), 1-9.

Rigby, R., Stasinopoulos, D., Akantziliotou, C., 2008. A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics & Data Analysis* 53(2), 381-393.

Rizzo, M.L., 2007. *Statistical computing with R*. CRC Press.

Saltzman, G.M., Belzer, M.H., 2007. Truck driver occupational safety and health: 2003 conference report and selective literature review.

Sanchez, R.O., Flores, C., Horowitz, R., Rajagopal, R., Varaiya, P., 2011. Arterial travel time estimation based on vehicle re-identification using magnetic sensors: Performance analysis, *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. IEEE, pp. 997-1002.

Schall, R., 1991. Estimation in generalized linear models with random effects. *Biometrika* 78(4), 719-727.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11), 2673-2681.

Searle, S., Casella, G., Mcculloch. CE (1992),“Variance Components”. New York: John Wiley & Sons.

Searle, S.R., Gruber, M.H., 2016. *Linear models*. John Wiley & Sons.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of Safety Research* 27(3), 183-194.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention* 29(6), 829-837.

Shefer, D., Rietveld, P., 1997. Congestion and safety on highways: towards an analytical model. *Urban Studies* 34(4), 679-692.

Shumway, R.H., Stoffer, D.S., 2010. *Time series analysis and its applications: with R examples*. Springer Science & Business Media.

Sivaraman, S., Trivedi, M.M., 2013. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems* 14(4), 1773-1795.

Speed, F.M., Hocking, R.R., Hackney, O., 1978. Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association* 73(361), 105-112.

Talebpour, A., Mahmassani, H., Hamdar, S., 2013. Speed harmonization: evaluation of effectiveness under congested conditions. *Transportation Research Record: Journal of the Transportation Research Board*(2391), 69-79.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society* 66(8), 1352-1362.

Ulfarsson, G.F., Mannering, F.L., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis & Prevention* 36(2), 135-147.

University of Arizona, U.o.C.P.P., Savari Networks Inc., Econolite, 2016. Multi-Modal Intelligent Traffic Signal System – Phase II: System Development, Deployment and Field Test Center for Transportation Studies.

USDOT, 2003. General estimates system coding and editing manual, *National Highway Traffic Safety Administration* Washington, DC.

USDOT, 2005. Report to Congress on the Large Truck Crash Causation Study, *Federal Motor Carrier Safety Administration*, Washington, DC.

USDOT, 2007. Corridors of the Future, *Federal Highway Administration*, Washington, DC.

Vapnik, V., 2013. *The nature of statistical learning theory*. Springer science & business media.

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Exploratory insights from random parameter negative binomial approach. *Transportation research record: journal of the transportation research board*(2236), 41-48.

WALLER, P., 2003. DEDICATED TO THE MEMORY OF. *OCCUPATIONAL SAFETY AND HEALTH*.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention* 41(4), 798-808.

Wang, S., Cui, L., Liu, D., Huck, R., Verma, P., Sluss, J.J., Cheng, S., 2012. Vehicle identification via sparse representation. *IEEE Transactions on Intelligent Transportation Systems* 13(2), 955-962.

WeatherUnderground, 2015. Arizona Weather Data, https://www.wunderground.com/.

Wolfinger, R., O'connell, M., 1993. Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation* 48(3-4), 233-243.

Wu, Q., Chen, F., Zhang, G., Liu, X.C., Wang, H., Bogus, S.M., 2014. Mixed logit model-based driver injury severity investigations in single-and multi-vehicle crashes on rural two-lane highways. *Accident Analysis & Prevention* 72, 105-115.

Xian, Y., Schiele, B., Akata, Z., 2017. Zero-Shot Learning-The Good, the Bad and the Ugly. *arXiv preprint arXiv:1703.04394*.

Xie, Y., Zhang, Y., Liang, F., 2009. Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering* 135(1), 18-25.

Zeng, N., Crisman, J.D., 1997. Vehicle matching using color, *Intelligent Transportation System, 1997. ITSC'97., IEEE Conference on*. IEEE, pp. 206-211.

Zha, L., Lord, D., Zou, Y., 2016. The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. *Journal of Transportation Safety & Security* 8(1), 18-35.

Zhang, R., Cao, L., Bao, S., Tan, J., 2017. A method for connected vehicle trajectory prediction and collision warning algorithm based on V2V communication. *International Journal of Crashworthiness* 22(1), 15-25.

Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention* 42(2), 626-636.

Zhou, M., Sisiopiku, V., 1997. Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record: Journal of the Transportation Research Board*(1581), 47-52.

Zhu, X., Srinivasan, S., 2011a. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accident Analysis & Prevention* 43(1), 49-57.

Zhu, X., Srinivasan, S., 2011b. Modeling occupant-level injury severity: An application to large-truck crashes. *Accident Analysis & Prevention* 43(4), 1427-1437.

Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476), 1418-1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 301-320.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics* 37(4), 1733.