8-18-2016

Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models

Kai Wang
University of Connecticut, kai.wang@uconn.edu

Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models

Kai Wang, Ph.D.

University of Connecticut, 2016

This report first describes the use of different copula based models to simultaneously estimate the two crash indicators: injury severity and vehicle damage. The Gaussian copula model outperforms the other copula based model specifications (*i.e.* Gaussian, Farlie-Gumbel-Morgenstern (FGM), Frank, Clayton, Joe and Gumbel copula models), and the results indicate that injury severity and vehicle damage are highly correlated, and the correlations between injury severity and vehicle damage varied with different crash characteristics including manners of collision and collision types. This study indicates that the copula-based model can be considered to get a more accurate model structure when simultaneously estimating injury severity and vehicle damage in crash severity analyses.

The second part of this report describes estimation of cluster based SPFs for local road intersections and segments in Connecticut using socio-economic and network topological data instead of traffic counts as exposure. The number of intersections and the total local roadway length were appropriate to be used as exposure in the intersection and segment SPFs, respectively. Models including total population, retail and non-retail employment and average household income are found to be the best both on the basis of model fit and out of sample prediction.

The third part of this report describes estimation of crashes by both crash type and crash severity on rural two-lane highways, using the Multivariate Poisson Lognormal (MVPLN) model. The crash type and crash severity counts are significantly correlated; the standard errors of covariates in the MVPLN model are slightly lower than the other two univariate crash prediction models (*i.e.* Negative Binomial model and Univariate Poisson Lognormal model) when the covariates are statistically significant; and the MVPLN model outperforms the UPLN and NB models in crash count prediction accuracy. This study indicates that when simultaneously predicting crash counts by crash type and crash severity for rural two-lane highways, the MVPLN model should be considered to avoid estimation error and to account for the potential correlations among crash type counts and crash severity counts.

Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models

Kai Wang

B.A., Xiamen University of Technology, 2011

M.A., South Dakota State University, 2014

A Dissertation

Submitted in Partial Fulfillment of the

Requirement of the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Exploration of Advances in Statistical Methodologies for Crash Count and Severity Prediction Models

Presented by

Kai Wang, B.A., M.A.

Major Advisor: _____

John N. Ivan

Associate Advisor: _____

Nalini Ravishanker

Associate Advisor: _____

Karthik C. Konduri

Associate Advisor: _____

Amy C. Burnicki

University of Connecticut

2016

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my major advisor Dr. John N. Ivan, for his patience, wisdom, enthusiasm and constant guidance and effort in past two years. I have learned a lot from his religious attitude toward research and work. Dr. Ivan taught me a lot both for my studies and lives. It is really my glories to be his student.

I would also like to thank Dr. Nalini Ravishanker, Dr. Karthik C. Konduri, Dr. Amy C. Burnicki and Dr. Norman W. Garrick, who are my associate advisors, for their suggestions on my research and comments on my dissertation. Thanks also go to Dr. Eric D. Jackson for his valuable ideas, advice and guidance on my research. Of course, I am always grateful to my Master advisor Dr. Xiao Qin from the University of Wisconsin - Milwaukee, for his guidance on my research and suggestions and help on my career. Thanks Dr. Qin for offering me the opportunity to study in the United States of America and accept the best education in the world.

Also, thanks the Connecticut Transportation Institute, and the department of Civil & Environmental Engineering for sponsoring my research. Thanks all my friends, fellows in transportation group and particularly my roommates Jie Qi, Shuai Zhao and Cheng Tu, who sharing my happiness and sorrow, and thanks my companions Zhiyuan Ma, Yihong Ning and Tao Gong who always enjoy swimming in the DOTA2 fishpond with me.

Last but not the least, I would like to thank my parents Tongjiang Wang and Zhixia Liu for giving me life, raising me and supporting me lifetime, and specially thank my girlfriend Qian Zhou for all her love, understanding and support, which always cheers me up every day.

TABLE OF CONTENTS

LIST OF FIGURES v
LIST OF TABLES vi
1 INTRODUCTION 1
1.1 BACKGROUND 1
1.2 OBJECTIVES AND METHODOLOGIES 2
1.3 REFERENCES 4
2 A COPULA BASED JOINT MODEL OF INJURY SEVERITY AND VEHICLE DAMAGE IN TWO-VEHICLE CRASHES 5
2.1 BACKGROUND 5
2.2 COPULA BASED MODEL 10
2.2.1 INJURY SEVERITY MODEL COMPONENT 10
2.2.2 VEHICLE DAMAGE MODEL COMPONENT 11
2.2.3 JOINT MODEL: A COPULA BASED APPROACH 11
2.3 DATA COLLECTION AND ANALYSIS 13
2.4 MODEL SPECIFICATIONS AND ASSUMPTIONS 17
2.5 MODEL ESTIMATION RESULTS 17
2.5.1 COEFFICIENT ESTIMATES 17
2.5.2 ELASTICITY EFFECTS 23
2.6 SUMMARY AND CONCLUSIONS 27
2.7 ACKNOWLEDGMENTS 29
2.8 REFERENCES 29
2.9 APPENDIX: PARAMETER ESTIAMTION AND EXPLANATION 36
3 PREDICTING LOCAL ROAD CRASHES USING SOCIO-ECONOMIC AND LAND COVER DATA 38
3.1 INTRODUCTION 38
3.2 LITERATURE REVIEW 40
3.3 METHODOLOGY AND DATA PREPARATION 43
3.3.1 ROADWAY NETWORK SHAPE FEATURES 43
3.3.2 TAZ LEVEL DEMOGRAPHIC RECORDS 44
3.3.3 TAZ LEVEL GEOGRAPHIC/LAND COVER FEATURES 44

3.3.4 CRASH RECORDS AND INTEGRATION OF CRASH TO TAZ 45
3.3.5 CLUSTERING OF TAZs 45
3.3.6 STATISTICAL METHODOLOGY 50
3.4 VARIABLE SELECTION AND SPF RESULTS 52
3.5 APPLICATIONS FOR NETWORK SCREENING 60
3.6 CONCLUSIONS AND FUTURE RESEARCH 60
3.7 ACKNOWLEDGMENTS 63
3.8 REFERENCES 63
4 MULTIVARIATE POISSON LOGNORMAL MODELING OF CRASHES BY TYPE AND SEVERITY ON RURAL TWO LANE HIGHWAYS 69
4.1 INTRODUCTION AND MOTIVATION 69
4.2 METHODOLOGIES 72
4.2.1 FRAMEWORK AND ESTIMATION FOR MVPLN MODEL 72
4.2.2 FRAMEWORK AND ESTIMATION FOR THE UPLN MODEL 74
4.2.3 FRAMEWORK AND ESTIMATION FOR NB Model 75
4.3 DATA PREPARATION AND ANALYSES 76
4.4 DISCUSSION OF RESULTS 80
4.5 MODEL COMPARISON 92
4.6 DISCUSSION AND CONCLUSION 97
4.7 ACKNOWLEDGMENTS 99
4.8 REFERENCES 99
4.9 APPENDIX: PARAMETER ESTIMATION AND EXPLANATION 103
5 CONCLUSIONS AND CONTRIBUTIONS 107

LIST OF FIGURES

FIGURE 3.1 Clustering Results and Cluster Distribution 47
FIGURE 3.2 Distributions of KAB Crashes by Cluster 47
FIGURE 3.3 Distributions of Independent Variables by Cluster 49
FIGURE 3.4 Distributions of Independent Variables by Cluster (Continued) 50
FIGURE 4.1 Prediction Performance of Crash Type Models (MN and WA Data) 94
FIGURE 4.2 Prediction Performance of Crash Type Models (CT Data) 94
FIGURE 4.3 Prediction Performance of Crash Severity Models (MN and WA Data) 95
FIGURE 4.4 Prediction Performance of Crash Severity Models (CT Data) 96

LIST OF TABLES

TABLE 2.1 Description of Selected Variables 15
TABLE 2.2 Estimated Results and Model Performances 18
TABLE 2.3 Gaussian Copula Model Coefficient Estimates and Copula Parameters 21
TABLE 2.4 Elasticity Effects for Vehicle Damage and Injury Severity 25
TABLE 3.1 Goodness-of-fit of the Cluster Based and Statewide SPFs 54
TABLE 3.2 SPF Prediction Performance 55
TABLE 3.3 Coefficient Estimates for KAB Intersection Crashes 57
TABLE 3.4 Coefficient Estimates for KAB Segment Crashes 59
TABLE 4.1 Descriptive Characteristics of 3ST Intersection Data 78
TABLE 4.2 Descriptive Characteristics of 4ST Intersection Data 78
TABLE 4.3 Descriptive Characteristics of 4SG Intersection Data 79
TABLE 4.4 Descriptive Characteristics of Segment Data 79
TABLE 4.5 Estimated Crash Type Models for MN 3ST Intersections (N=755 Intersections) 82
TABLE 4.6 Estimated Crash Severity Models for MN 3ST Intersections (N=755 Intersections) 83
TABLE 4.7 Estimated Crash Type Models for MN 4ST Intersections (N=1064 Intersections) 85
TABLE 4.8 Estimated Crash Severity Models for MN 4ST Intersections (N=1064 Intersections) 86
TABLE 4.9 Estimated Crash Type Models for MN 4SG Intersections (N=63 Intersections) 88
TABLE 4.10 Estimated Crash Severity Models for MN 4SG Intersections (N=63 Intersections) 89
TABLE 4.11 Estimated Crash Type Models for WA Rural Two-Lane Segments (N=7583 Segments) 91
TABLE 4.12 Estimated Crash Severity Models for WA Rural Two-Lane Segments (N=7583 Segments) 92

1 INTRODUCTION

1.1 BACKGROUND

Improving traffic safety was, is and will continue to be a high priority on the national transportation agenda due to the significant social and financial implications of motor vehicle crashes including injuries, deaths and economic losses, among others. In the past few decades, organizations such as Federal Highway Administration (FHWA) (1), National Highway Traffic Safety Administration (NHTSA) (2), and American Association of State Highway and Transportation Officials (AASHTO) (3) have launched numerous roadway safety campaigns and implemented various strategies for reducing the number of crashes with a particular emphasis on reducing the most severe ones (4). These efforts have been targeted at different aspects of roadway safety from improvements in highway engineering, to driver education, to driver assistance technologies and traffic enforcements. All of these efforts have led to a significant reduction in traffic fatalities, from 43,510 in 2005 to about 32,675 in 2014 (a 25 percent reduction in 10 year span) (2). However, traffic safety still remains a significant issue and more needs to be done to alleviate the negative implications of crashes. In order to implement effective safety strategies and countermeasures, it is necessary to identify the different factors contributing to crashes and factors affecting crash severity in the event of a crash.

There are two critical aspects of traffic safety analysis addressed in this research: crash counts and crash severity. Crash counts usually represent the number of crashes for a roadway facility (*e.g.* rural highways, urban and suburban highways and freeways) in a specific time period. Crash severity usually represents the highest injury level in a crash, which is categorized into five types: type K (fatal injury), type A (suspected serious injury), type B (suspected minor

injury), type C (possible injury) and type O or PDO (no apparent injury or property damage only). In order to implement highway safety improvement strategies to reduce crashes and crash severity, appropriate statistical approaches are desirable to predict the crash counts or crash severity, as well as exploring the contributors affecting the crash counts and crash severity.

1.2 OBJECTIVES AND METHODOLOGIES

In this report, three distinct objectives are considered using three statistical models.

1. The first objective of this research is to simultaneously model the two indicators of crash severity - injury severity and vehicle damage, and to identify the contributing factors, as well as to explore the potential correlation between the two indicators across crashes due to common unobserved attributes. A copula based model is applied as this approach can jointly model two dependent variables and account for the correlation between them through a copula structure.
2. The second objective of this research is to estimate Safety Performance Functions (SPFs) for both intersection and segments on roads under local jurisdiction in the State of Connecticut where the AADT is not available. The SPFs are estimated at the level of Traffic Analysis Zone (TAZ), using socio-economic data and network topological data as a replacement for traffic count data. To account for data and crash relationship heterogeneity, the TAZs are categorized into different clusters based on the percentage of three land cover categories – high, medium and low intensities – and the population density (*i.e.* the number of population per km^2) to account for differences in the crash occurrence phenomenon among these different area types.
3. The third objective of this research is to simultaneously estimate crashes by both crash type and crash severity on rural two-lane highways, and explore the possible correlations among

crash type or crash severity counts. The Multivariate Poisson Lognormal (MVPLN) model was developed and the approximate Bayesian inference via the Integrated Nested Laplace Approximation (INLA) was implemented. This approach can simultaneously model crash counts by crash type and crash severity by accounting for the potential correlations among them and significantly decreases the computational time compared with a fully Bayesian fitting of the MVPLN model using Markov Chain Monte Carlo (MCMC) methods. The MVPLN models were developed for three-way stop controlled (3ST) intersections, four-way stop controlled (4ST) intersections, four-way signalized (4SG) intersections, and roadway segments on rural two-lane highways. Annual Average Daily traffic (AADT) and variables describing roadway conditions (including presence of lighting, presence of left-turn/right-turn lane, lane width and shoulder width) were used as predictors.

This research contains three published or under-reviewed journal papers (the first paper was published at the Journal of Transportation Research Record; the second paper was presented at the 2016 annual meeting of Transportation Research Board and published at the Journal of Transportation Safety & Security; the third paper was submitted to Journal of Accident Analysis & Prevention for publication and is currently under review), and contributes substantively to the exploration of methodologies in traffic safety analysis. We propose separate models for crash counts and severity analysis, with unbiased variance estimates for coefficients. Contributing factors associated with crash counts and crash severities are identified and relevant countermeasures are recommended to reduce the effects of crashes. The rest of the report is organized as follows. The next section presents a copula based methodology in crash severity analysis. The third section describes an approach for predicting crashes on local roads using socio-economic variables rather than AADT. The fourth section demonstrates the exploration of

a Multivariate Poisson Lognormal Model in predicting crashes by crash type and severity on rural two lane highways. The conclusions and contributions are summarized in the final section.

1.3 REFERENCES

1. U.S. Department of Transportation. Federal Highway Administration (FHWA). http://www.fhwa.dot.gov/
2. Unites States. NHTSA National Center for Statistic and Analysis. Passenger Vehicle Occupant Fatalities: The Decline for Six Years in a Row from 2005 to 2011. http://www-nrd.nhtsa.dot.gov/Pubs/812160.pdf. 2015.
3. American Association of State Highway and Transportation Official (AASHTO). http://www.transportation.org/Pages/Default.aspx
4. FHWA-SA-10-005. Federal Highway Administration (FHWA). The U.S. Department of Transportation. 2009. http://safety.fhwa.dot.gov/intersection/resources/fhwasa10005/docs/brief_2.pdf

2 A COPULA BASED JOINT MODEL OF INJURY SEVERITY AND VEHICLE DAMAGE IN TWO-VEHICLE CRASHES

2.1 BACKGROUND

Improving traffic safety was, is and will continue to be a high priority on the national transportation agenda due to the significant social and financial implications of motor vehicle crashes including injuries, deaths and economic losses among others. In the past few decades, organizations such as Federal Highway Administration (FHWA) (1), American Association of State Highway and Transportation Official (AASHTO) (3) have launched numerous roadway safety campaigns and implemented various strategies for reducing the number of crashes with a particular emphasis on reducing the severe ones (4). These efforts have been targeted at different aspects of roadway safety from improvements in highway engineering, to driver education, to driver assistance technologies and traffic enforcements. All of these efforts have led to a significant reduction in traffic fatalities from 43,510 in 2005 to about 32,367 in 2011(a 26 percent reduction in 7 year span) (2). However, traffic safety still remains a significant externality and more needs to be done to alleviate the negative implications of crashes. In order to implement effective safety strategies and countermeasures, it is necessary to identify the different factors contributing to crashes and factors affecting crash severity in the event of a crash.

Injury severity is an important indicator that is usually modeled to identify the different factors contributing to driver injuries. Discrete choice methodologies have commonly been used to model the effects of driver, environmental, highway, traffic, and vehicle factors on injury severity (5, 6). Among the different discrete choice methodologies, logistic and probit model

formulations have been extensively used to examine the relationship between the contributing factors and injury severity. In studies where injury severity is treated as a non-ordinal indicator, the multinomial logistic or probit model formulations have been used to investigate the relationship between contributing factors and injury severity (7-10). In studies where injury severity is treated as an ordinal variable, traditional ordered logistic or probit model formulations or generalized ordered logit formulations have been used (11-21).

Both ordered and unordered logistic or probit models are fixed parameter models in which all parameters are assumed to be constant across observations. However, it is argued that model coefficients may not remain constant but vary across individuals when the data are heterogeneous. To this end, other model formulations were proposed to capture the heterogeneity across crashes. The Markov switching multinomial logistic model was used to account for unobserved factors that influence injury severity (22). The random parameter (mixed) model is an alternative formulation which can treat the parameters as either fixed or random variables (7, 9, 20, 23-27). More recently latent segmentation models that account for heterogeneity in a closed form structure in severity models have also been employed (18). Savolainen et al. (28) reviewed and summarized numerous discrete choice models that are currently being used in modeling injury severity and offered additional insights about model evaluation and selection.

Recently, in order to capture the interrelationships among variables when the factors interact in indirect and complicated ways in injury severity models, researchers have also extensively applied the structural equation modeling (SEM) in exploring the contribution of different explanatory variables on injury severity. SEM can effectively account for complex relationships

between multiple dependent and independent variables simultaneously. Further, SEM can also incorporate the influence of latent variables on dependent variables of interest (5, 6, 29-34).

Although injury severity has been used extensively in modeling the severity of a crash, it may not be the most representative indicator. Injury severity is a subjective indicator based on victim's responses, descriptions, and complaints after the crash; owing to the self-reported nature of the measure, it may be prone to bias (6). On the other hand, the extent of vehicle damage is a more objective indicator based on the destruction/deformation of the vehicle involved in the crash; as it can be readily seen and measured. Due to its objective nature, vehicle damage has been used as an additional indicator to characterize crash severity (6, 35-37).

Although vehicle damage has been introduced as an additional indicator in crash severity analysis, the treatment and modeling of the different indicators is up for debate. Injury severity and vehicle damage are typically modeled independently which may lead to possible estimation and inference issues because the two indicators are likely to be correlated (35). The levels of the indicators for any given crashes are correlated due to both observed and unobserved factors. Although the correlations due to the observed factors can be accounted for by specifying them as explanatory variables, same cannot be said about the unobserved factors because they are not observable. Ignoring the correlations due to unobserved factors may result in incorrect and biased coefficient estimates (38). Therefore, there is a need for model formulations that can simultaneously model the injury severity and vehicle damage indicators of crash severity while also accounting for potential interrelationships between the indicators.

In this study, the copula based approach is used to model the injury severity and vehicle damage dimensions simultaneously while also accounting for the error correlations that may exist across the two dimensions. Further, in the copula approach, parameterization of the copula structure is allowed to help explain the heterogeneity in correlations between the dependent variables (39). In recent years, the copula based model has been increasingly used in transportation research.

Pourabdollahi *et al.* (40*)* used a copula based model to estimate the choice of freight mode and shipment size simultaneously. The study confirms that the copula based model can effectively capture the effects of common unobserved factors affecting both variables, and consequently it can appropriately account for the correlations between the selection of freight mode and shipment size. Sener *et al.* (41*)* applied a copula based model to examine the physical activity participation for all individuals within the same family unit, by accounting for the dependencies among individuals' activity participation due to the common observed and unobserved factors. The model results show that individuals in the same family unit tend to have simultaneously low physical activity propensities, while the trend for high propensities is not significant.

The copula based model has also been used in modeling crash severity. Eluru *et al.* (42) examined the injury severities for all occupants involved in a crash using a copula based model. The effects of common unobserved factors on all occupants in the same vehicle were accommodated in the model. The results illustrate that the copula based model is better than the independent ordered probit model (in which the injury severity for each occupant was independently and separately modeled) with regard to the model goodness-of-fit. The study conducted by Rana *et al.* (43) employed a copula based model to consider the crash type and

injury severity as dependent variables simultaneously. The model estimation results show that the copula based model outperforms the independent models in which the collision type and injury severity were independently modeled. Yasmin *et al.* (39) improved the model developed by Rana *et al.* by allowing the dependencies between injury severity and collision type to vary across different categories of collision type. The results suggest that injury severity and collision type are correlated, and the correlation between injury severity and collision type varies with the type of collision.

The research presented here is an attempt to model the injury severity and vehicle damage and to identify contributing factors, while also accounting for the potential correlations between the two indicators due to unobserved attributes. To this end, the copula based approach is applied to simultaneously model injury severity and vehicle damage. Given the ordered nature of the injury severity and vehicle damage indicators, ordered probit formulation was used to model both of the two indicators. The error correlations between the injury severity and vehicle were tied together using different copula formulations and parameterization strategies. The proposed model is estimated using the five-year (2005-2009) crash data for two-vehicle crashes collected from the Madison, Wisconsin, including a detailed set of exogenous variables, *i.e.*, driver characteristics, highway and traffic factors, environmental factors and crash characteristics. The rest of the paper is organized as follows. The next section presents the copula based methodology used in this paper. The third section describes the data in detail and the fourth section presents the model specifications and assumptions. The model results are presented in the fifth section, and concluding thoughts are presented in the final section.

2.2 COPULA BASED MODEL

The primary objective of this study is to simultaneously model the injury severity and vehicle damage levels of crashes using a copula based approach. The indicators are treated as ordinal variables and a probit formulation is used to model the indicators. The econometric formulation of the proposed copula methodology is presented below:

2.2.1 INJURY SEVERITY MODEL COMPONENT

Let q ($q = 1,2, \dots, Q$) be the index for vehicle involved in the crash, j ($j = 1,2, \dots, J$) be the index representing the level of injury severity and k ($k = 1,2, \dots, K$) be the index representing the level of vehicle damage. In an ordered probit formulation, the discrete injury severity level (y_q) is assumed to be associated with an underlying continuous latent propensity (y_q^*). Further, the latent propensity is specified as follows:

$$y_q^* = \alpha' x_q + \varepsilon_q \,, \;\; y_q = j, if \; \tau_{j-1} < y_{qj}^* < \tau_j \qquad (2\text{-}1)$$

where, y_q^* is the latent propensity of injury severity for vehicle q, x_q is a vector of exogenous variables, α is the associated row vector of unknown parameters and ε_q is a random disturbance term assumed to be standard normal. τ_j ($\tau_0 = -\infty \,, \tau_J = \infty$) represents the threshold associated with severity level j, with the following ordering conditions: $\left(-\infty < \tau_1 < \tau_2 < \; \dots < \tau_{J-1} < +\infty\right)$. Given the above information regarding the different parameters, the resulting probability expression for the occupant of vehicle q sustaining an injury severity level j takes the following form:

$$Pr(y_q = j) = \phi\left(\tau_j - \alpha' x_q\right) - \phi\left(\tau_{j-1} - \alpha' x_q\right) \qquad (2\text{-}2)$$

where, $\phi(\cdot)$ is the cumulative standard normal distribution function. The probability expression in Equation (2-2) represents the independent injury severity model for the occupant of vehicle q.

2.2.2 VEHICLE DAMAGE MODEL COMPONENT

On the other hand, vehicle damage component also takes the form of an ordered probit formulation. The expression for latent propensity (u_q^*) of vehicle damage is shown below:

$$u_q^* = \beta' z_q + \xi_q\,,\ \ u_q = k, if\ \psi_{k-1} < u_{qk}^* < \psi_k \tag{2-3}$$

where, u_q^* is the latent propensity of vehicle damage for vehicle q, u_q is the discrete level of vehicle damage, z_q is a vector of exogenous variables, β is the associated row vector of unknown parameters, ξ_q is a random disturbance term assumed to be standard normal and ψ_k represents the threshold associated with vehicular damage level k. Assuming similar information for the thresholds as in the injury severity model component, the probability expressions for vehicle q with a damage level k can be written as:

$$Pr(u_q = k) = \Lambda\left(\psi_k - \beta' z_q\right) - \Lambda\left(\psi_{k-1} - \beta' z_q\right) \tag{2-4}$$

where, $\Lambda(\cdot)$ is the cumulative standard normal distribution function.

2.2.3 JOINT MODEL: A COPULA BASED APPROACH

In examining the injury severity and vehicle damage simultaneously, the dependency between the two dimensions of interests is captured through the error terms (ε_q and ξ_q) from equation (2-1) and (2-3). The joint probability of sustaining injury severity level *j* and vehicle damage level k for vehicle q can be expressed as:

$$\begin{aligned} &Pr\left(y_q = j, u_q = k\right) \\ &\quad = Pr\left[\left(\left(\tau_{j-1} - \alpha' x_q\right) < \varepsilon_q < \left(\tau_j - \alpha' x_q\right)\right), \left((\psi_{k-1} - \beta' z_q) < \right.\right. \\ &\quad\quad \left.\left. \xi_q < (\psi_k - \beta' z_q)\right)\right] \\ &\quad = Pr\left[\varepsilon_q < \left(\tau_j - \alpha' x_q\right),\ \xi_q < (\psi_k - \beta' z_q)\right] \end{aligned} \tag{2-5}$$

$$-Pr\left[\varepsilon_q < \left(\tau_j - \alpha' x_q\right),\ \xi_q < (\psi_{k-1} - \beta' z_q)\right]$$

$$-Pr\left[\varepsilon_q < \left(\tau_{j-1} - \alpha' x_q\right),\ \xi_q < (\psi_k - \beta' z_q)\right]$$

$$+Pr\left[\varepsilon_q < \left(\tau_{j-1} - \alpha' x_q\right),\ \xi_q < (\psi_{k-1} - \beta' z_q)\right]$$

Given the above setup, the correlations between the injury severity and vehicle damage due to unobserved factors are accommodated using a copula based approach. A detailed description of the copula approach can be found in Bhat and Eluru (44), Trivedi and Zimmer (45). The joint probability of equation (2-5) can be expressed by using the copula function as:

$$Pr\left(y_{qj} = j, u_{qk} = k\right) = C_{\theta q}\left(U_{qj}, U_{qk}\right) - C_{\theta q}\left(U_{qj}, U_{qk-1}\right) - C_{\theta q}\left(U_{qj-1}, U_{qk}\right) + C_{\theta q}\left(U_{qj-1}, U_{qk-1}\right) \quad (2\text{-}6)$$

It is important to note here that the level of dependence between injury severity level and vehicle damage can vary across crashes. Therefore, in the current study, the dependence parameter θ_q is parameterized as a function of observed crash attributes as follows:

$$\theta_q = f_n(\gamma' s_q) \quad (2\text{-}7)$$

where, s_q is a column vector of exogenous variables, γ' is the associated row vector of unknown parameters (including a constant) and f_n represents the functional form of parameterization. In this study, six different copula structures are respectively explored: Gaussian, Farlie-Gumbel-Morgenstern (FGM), Frank, Clayton, Joe and Gumbel copulas. A detailed discussion of these copulas is available in Bhat and Eluru (44). Based on the permissible ranges of the dependency parameter, different functional forms are assumed for the parameterization of the six copula structures in the analysis. For Gaussian and Farlie-Gumbel-Morgenstern (FGM) copulas, functional form $\theta_q = \gamma' s_q$ is used. For the Clayton and Frank copulas, $\theta_q = exp(\gamma' s_q)$ is applied. Finally for Joe and Gumbel copulas, $\theta_q = 1 + exp(\gamma' s_q)$ is assumed. Further, similar parameterizations can be found in Sener *et al.* (41), Eluru *et al.* (42) and Yasmin *et al.* (39).

Of the six copulas, Clayton, Joe and Gumbel allow for asymmetric copulas that consider dependency in one direction. To potentially account for the possibility of a reverse dependency, with asymmetric copulas, a reverse dependent variable was considered for vehicle damage (wherein a new dependent variable is created by sorting vehicle damage from highest level to lowest level). This reversing of the dependent variables does not affect the ordered probit model probabilities (except for changes to the threshold values).

With the above as preliminaries, the likelihood function can be expressed as:

$$L = \prod_{q=1}^{Q} \left[\prod_{j=1}^{J} \prod_{k=1}^{K} \{Pr(y_q = j, u_q = k)\}^{\omega_{qkj}} \right] \quad (2\text{-}8)$$

where, ω_{qkj} is a dummy indicator variable assuming a value of 1 if injury severity level is j and vehicle damage level is k for the vehicle q and 0 otherwise. All the parameters in the model are consistently estimated by maximizing the logarithmic function of L. The parameters to be estimated in the model are: α' and τ_j in the injury severity component, β' and ψ_k in vehicle damage component, and finally γ' in the dependency component.

2.3 DATA COLLECTION AND ANALYSIS

In this study, crash data collected in Madison, Wisconsin between 2005 and 2009 was used and only two-vehicle crashes were considered. Between 2005 and 2009, there were 13,683 two-vehicle crashes in Madison, Wisconsin, accounting for 60 percent of all crashes. Among all two-vehicle crashes, according to the Model Minimum Uniform Crash Criteria (MMUCC) guideline or "KABCO" scale (46), 9,488 or 69.3 percent crashes were type O (no apparent injury or property damage only); 4,062 or 29.7 percent crashes were either type B (suspected minor injury) or C (possible injury); and 133 or 1 percent crashes were either type A (suspected serious injury)

or K (fatal injury). With regard to the vehicle damage, referring to the Wisconsin Motor Vehicle Report Form (MV 4000) (47), 4,640 or 33.9 percent were none (no damage) or minor (cosmetic damage); 6,250 or 45.7 percent were moderate (broken or missing parts); and 2,793 or 20.4 percent were severe (salvageable) or very severe (total loss).

Factors contributing to crashes in the database were categorized into four groups: driver characteristics, highway and traffic factors, environmental factors and crash characteristics. Driver characteristics include driver's age, gender, usage of safety restraints and whether the driver was driving under the influence of alcohol or drugs. Highway and traffic factors include the highway geometric characteristics, highway class and traffic control types. Environmental factors include weather, light and roadway surface conditions. Crash characteristics include the manner of collision which describes the orientation that vehicles collided, and the collision type which indicates the types of vehicles that collided with each other. The detailed description of selected variables is shown in Table 2.1.

TABLE 2.1 Description of Selected Variables

Category	Variable	Type and Value	Description	Frequency	Percentage
Driver	AGE	**Categorical**	Driver age		
Characteristics		1	Young (<25)	3,805	27.8%
		2	Middle (25-55)	7,688	56.2%
		3	Old (>55)	2,190	16.0%
	GENDER	**Dummy**	Male driver	7,047	51.5%
	DUI	**Dummy**	Driver under the influence of drugs or alcohol	524	3.8%
	SAFETY	**Dummy**	Safety restraints	13,323	97.4%
Highway and	ROADHOR	**Dummy**	Horizontal curve	1,045	7.6%
Traffic Factors	ROADVERT	**Dummy**	Vertical curve	1,826	13.3%
	HWYCLASS	**Categorical**	Highway class		
		1	Urban city highway	9,909	72.4%
		2	Urban state highway	3,549	25.9%
		3	Urban interstate highway	225	1.7%
	TRFCONT	**Categorical**	Traffic control		
		1	Four-way stop sign (intersection)	344	2.5%
		2	Two-way stop sign (intersection)	1,491	10.9%
		3	Signal (intersection)	4,478	32.7%
		4	Yield or no control (intersection)	1,988	14.5%
		5	No control (segment)	5,382	39.4%
Environmental	WTHRCOND	**Categorical**	Weather condition		
Factors		1	Clear	7,290	53.3%
		2	Cloudy	4,118	30.1%
		3	Rain	1,289	9.4%
		4	Snow/hail	986	7.2%
	LGTCOND	**Categorical**	Light condition		
		1	Day	10,059	73.5%
		2	Night without street light	941	6.9%
		3	Night with street light	2,683	19.6%
	ROADCOND	**Categorical**	Road surface condition		

		1	Dry	9,206	67.3%
		2	Wet	2,448	17.9%
		3	Snow/slush	1,495	10.9%
		4	Ice	534	3.9%
Crash	MNRCOLL	**Categorical**	Manner of collision		
Characteristics		1	Head-on	277	2.0%
		2	Rear-end	5,295	38.7%
		3	Sideswipe (same/opposite direction)	2,588	18.9%
		4	Angle	5,523	40.4%
	COLLTYPE	**Categorical**	Collision type		
		1	PC with PC	10,148	74.2%
		2	PC with truck	3,243	23.7%
		3	Truck with truck	292	2.1%
Crash Severities	INJSVR	**Ordinal**	Injury severity level		
		1	O	9,488	69.3%
		2	C+B	4,062	29.7%
		3	A+K	133	1.0%
	VEHDMG	**Ordinal**	Vehicle damage level		
		1	None or minor	4,640	33.9%
		2	Moderate	6,250	45.7%
		3	Severe or very severe	2,793	20.4%

2.4 MODEL SPECIFICATIONS AND ASSUMPTIONS

Using a copula-based model, injury severity and vehicle damage indicators were jointly modeled to explore factors contributing to the crash outcomes. The joint model contains an injury severity component and a vehicle damage component. In the injury severity component, all four categories of explanatory variables: driver characteristics, highway and traffic factors, environmental factors, and crash characteristics were explored. On the other hand, in the vehicle damage component, driver characteristics were not considered because it was assumed that vehicle damage is affected by highway and traffic factors, environmental factors, and crash characteristics. The detailed discussion and explanation of the variable selection for injury severity and vehicle damage models can be found in a previous study conducted by Qin *et al.* (35).

Six different copula structures were explored in this study: the Gaussian, FGM, Frank, Clayton, Joe and Gumbel copulas. The model development process comprised of the following three steps: 1) and independent model of injury severity and vehicle damage was estimated to serve as the starting point for the joint model estimation and also for purposes of comparison with the joint model, 2) copula models using the six different types of copulas were estimated, 3) and finally, the six copula models were compared with the independent model and with each other; Bayesian Information criterion (BIC) criterion was used to determine the best model *(42).*

2.5 MODEL ESTIMATION RESULTS

2.5.1 COEFFICIENT ESTIMATES

As noted earlier, six different copula models and an independent model were estimated in this study. The performance of the best five models is listed in Table 2. Based on the model

goodness-of-fit, all six copula based models have a lower BIC value than the independent model. This indicates the correlations caused by unobserved factors between injury severity and vehicle damage do exist, and accounting for these dependencies can improve model accuracy. The BIC metric for the independent model and best fitting four copula models are presented in Table 2.2. Among the copula based models, the model with a Gaussian copula structure was found to provide the lowest BIC value thereby indicating that the model best fits the data.

TABLE 2.2 Estimated Results and Model Performances

Models	Number of Estimated Parameters	BIC
Independent Model	24	44,862.79
Gaussian Copula Model	29	44,037.98
FGM Copula Model	26	44,415.52
Frank Copula Model	29	44,071.91
Clayton Copula Model	26	44,698.22

Table 2.3 presents the coefficient estimates of Gaussian copula based model for injury severity and vehicle damage. The table also presents the results of the copula structure parameterization. In the table, a positive value of a coefficient in the model of injury severity (vehicle damage) represents a propensity to increase the injury severity (vehicle damage) and vice-versa for a negative value of a coefficient. On the other hand, a positive value in the copula structure parameterization represents a positive correlation between the common unobserved factors affecting injury severity and vehicle damage and a negative coefficient represents a negative dependency between the common unobserved factors affecting injury severity and vehicle damage.

Driver related factors play an important role in any crash severity studies. It can be seen from Table 2.3 that all human factors have significant influences on injury severity outcomes. It was found that young drivers are less likely to relate to severe injuries compared with others. This is

possibly due to the higher physiological strength of younger drivers compared to elderly drivers (39). A negative coefficient was also estimated for male drivers. Consistent with expectation, compliance with law is highly associated with the slight injury severity. It was found that the use of alcohol or drugs considerably relates to the probability of severe injury severity while using safety restraints dramatically decrease the probability of severe severity of injury.

Highway and traffic factors are of interest to highway and traffic engineers for designing and implementing cost-effective countermeasures to improve highway safety. Based on the coefficient estimates of the highway class for injury severity and vehicle damage, it can be seen that crashes occurring on the interstate highway are the most severe ones, followed by those occurring on state and city highways. This is possibly due to higher speeds associated with interstate facilities compared to other highway functional classes (35). With regard to the traffic control types, four-way stop appears to be the safest traffic control strategy. Four way stop sign is less likely associated with severe injury severity compared to all other traffic controls at intersections and it is also less likely associated with severe severity of vehicle damage compared with all intersection traffic controls. This is plausible because four-way stop controlled intersections experience the smallest speed differentials between intersecting highways compared with others thereby leading to lower levels of injury severity and vehicle damage in the event of a crash (35).

Environmental factors were also found to affect both injury severity and vehicle damage. It is interesting to note that adverse roadway conditions are more likely to be associated with slight injury severity and slight vehicle damage. This is possibly due to the reduction in speeds by

drivers for cautionary reasons during adverse weather conditions (7). One of the most interesting finding is with regard to the lighting conditions. It was found that crashes caused at night time are related with severe vehicle damage irrespective of the street lighting conditions. However, no such influence was found on injury severity. This can be supported by the study conducted by Qin *et al.* (35) in which the authors concluded that the structural design of the vehicle can protect occupants from sustaining injuries, but severe collisions may reduce the effectiveness of the protection.

With regard to the manner of collision, compared with the rear-end crashes, head-on crashes are significantly associated with the severe injury severity; both head-on and angle crashes are associated with severe vehicle damage. For the collision type, crashes between two passenger cars are significantly associated with severe injury severity and vehicle damage compared with those between a passenger car and a truck as well as between two trucks. This is possible due to the larger speed differentials between two passenger cars.

TABLE 2.3 Gaussian Copula Model Coefficient Estimates and Copula Parameters

Gaussian Copula Ordered Probit-Ordered Probit Model									
Variable		**Injury Severity Component**				**Vehicle Damage Component**			
		Coef.	**SE**	**t**	**P > |t|**	**Coef.**	**SE**	**t**	**P > |t|**
Driver characteristics									
Age	Old	Base level				NA			
	Middle	---				NA			
	Young	-0.24	0.03	-9.67	<0.01	NA			
Gender	Male driver	-0.23	0.02	-10.28	<0.01	NA			
DUI	Drug or alcohol	0.31	0.05	6.03	<0.01	NA			
Safety	Safety restraints	-0.6	0.06	-9.87	<0.01	NA			
Highway and traffic factors									
Curve	Horizontal curve	---				---			
	Vertical curve	---				---			
Highway class	Urban city highway	Base level				Base level			
	Urban state highway	0.10	0.03	3.76	<0.01	0.06	0.02	2.68	0.01
	Urban interstate highway	0.18	0.09	2.15	0.03	0.35	0.07	4.78	<0.01
Traffic control	No control (segment)	Base level				Base level			
	Two-way stop sign (intersection)	0.13	0.04	3.50	<0.01	---			
	Signal (intersection)	0.13	0.03	5.09	<0.01	---			
	Yield or no control (intersection)	0.08	0.03	2.45	0.01	---			
	Four-way stop sign (intersection)	---				-0.32	0.06	-5.04	<0.01
Environmental factors									
Weather condition	Clear	Base level				Base level			
	Cloudy	---				---			
	Rain	---				---			
	Snow/hail	---				---			
Light condition	Day	Base level				Base level			

	Night without street light			---		0.09	0.04	2.40	0.02
	Night with street light			---		0.10	0.02	4.19	<0.01
Roadway condition	Dry			Base level				Base level	
	Wet			---				---	
	Snow/slush	-0.17	0.04	-4.66	<0.01	-0.13	0.03	-4.26	<0.01
	Ice	-0.17	0.06	-2.88	<0.01	-0.10	0.05	-1.93	0.05
Crash characteristics									
Manner of collision	Rear-end			Base level				Base level	
	Head-on	0.44	0.07	5.94	<0.01	0.97	0.07	14.52	<0.01
	Sideswipe (same/opposite direction)	-0.60	0.03	-18.44	<0.01			---	
	Angle			---		0.64	0.02	31.58	<0.01
Collision type	Truck with truck			Base level				Base level	
	Passenger car with truck			---				---	
	Passenger car with passenger car	0.08	0.03	2.97	<0.01	0.06	0.02	2.69	0.01
Threshold	μ1			-0.19				-0.10	
	μ2			1.74				1.22	

Copula Parameters	**Coef.**	**SE**	**t**	**P > |t|**
Constant	0.11	0.03	4.12	<0.01
Passenger car with passenger car	-0.11	0.02	-4.40	<0.01
Head-on	0.46	0.07	6.34	<0.01
Angle	0.45	0.02	18.06	<0.01
Sideswipe (same/opposite direction)	0.39	0.03	11.37	<0.01

Notes: “NA” represents “not applicable”; “---” represents the variable is not statistically at 5% level of significance.

The estimated copula parameters offered additional insight about the dependencies between injury severity and vehicle damage. In determining variables for the copula structure, we first select all candidate variables, and then remove variables that are not statistically significant. In Table 2.3, only the parameters for the copula structure that have been considered to be statistically significant at 5% level of significance are included.

The results highlight the existence of dependencies between injury severity and vehicle damage caused by the common unobserved factors. A positive parameter indicates that the dependencies between injury severity and vehicle damage caused by the common unobserved factors for the specific type of crashes are positive, and a negative parameter indicates that the dependencies between injury severity and vehicle damage caused by the common unobserved factors for the specific type of crashes are negative. It is interesting to note that the dependencies vary with different characteristics of crashes including manners of collision and collision types. With regard to three manners of collision: head-on, angle and sideswipe, the dependencies between injury severity and vehicle damage caused by the common unobserved factors were found to be positive. The magnitude of copula parameters implies that the highest level of dependency between injury severity and vehicle damage is for head-on crashes, followed by angle and sideswipe crashes. Also, the dependencies between injury severity and vehicle damage for crashes between two passenger cars were shown to be negative.

2.5.2 ELASTICITY EFFECTS

In the copula based model, the estimated parameters alone are not sufficient to describe the magnitude of the effect of an independent variable on the probability of each vehicle damage or injury severity category. Therefore, the elasticity effects for all independent variables with regard

to both injury severity and vehicle damage were calculated and are presented in Table 2.4. The detailed discussion on the methodology for calculating elasticity effects in a copula based model can be found in Eluru and Bhat (48).

In general, the effects of independent variables on injury severity and vehicle damage shown in Table 2.4 are consistent with those described in Table 2.3. More specifically, the presence of young and male drivers decreases the probability of severe injury severity, the use of drug or alcohol significantly increase the probability of severe injuries, and using safety restraints dramatically decreases the probability of severe injuries especially the type A or fatal injuries. With regard to highway and traffic factors, roadways with higher speed limit increase the probability of both severe injuries and vehicle damage levels. Four-way stop controlled intersections decrease the probability of severe crash outcomes. In terms of the environmental factors, adverse roadway surface conditions seem to decrease the probability of injury type B or C and type A and K, as well as decreasing the probability of moderate and severe vehicle damages. Night time with or without street lights increases the probability of severe vehicle damages, but the effects of it on injury severity were not statistically significant. The crash characteristics describe the manner and vehicle type of a collision. Head-on crashes have the most significant impacts on increasing severe crash severities, and collisions between two passenger cars are the most severe ones among all collision types.

TABLE 2.4 Elasticity Effects for Vehicle Damage and Injury Severity

Variable		Injury Severity			Vehicle Damage		
		PDO	C+B	A+K	None+ Minor	Moderate	Severe+ Very Severe
Driver characteristics							
Age	Old		Base level		NA	NA	NA
	Middle	---	---	---	NA	NA	NA
	Young	8.69	-12.81	-21.07	NA	NA	NA
Gender	Male driver	8.39	-12.19	-20.98	NA	NA	NA
DUI	Drug or alcohol	-11.74	16.26	32.16	NA	NA	NA
Safety	Safety restraints	22.67	-29.65	-68.30	NA	NA	NA
Highway and traffic factors							
Curve	Horizontal curve	---	---	---	---	---	---
	Vertical curve	---	---	---	---	---	---
Highway class	Urban city highway		Base level			Base level	
	Urban state highway	-3.50	5.05	8.88	-3.52	0.51	4.12
	Urban interstate highway	-6.84	9.65	18.11	-20.01	0.70	25.63
Traffic control	No control (segment)		Base level			Base level	
	Two-way stop sign (intersection)	-4.75	6.79	12.25	---	---	---
	Signal (intersection)	-4.94	7.12	12.51	---	---	---
	Yield or no control (intersection)	-3.05	4.39	7.78	---	---	---
	Four-way stop sign (intersection)	---	---	---	19.23	-4.78	-20.44
Environmental factors							
Weather condition	Clear		Base level			Base level	
	Cloudy	---	---	---	---	---	---
	Rain	---	---	---	---	---	---
	Snow/hail	---	---	---	---	---	---
Light condition	Day		Base level			Base level	
	Night without street light	---	---	---	-5.05	0.65	5.98
	Night with street light	---	---	---	-5.89	0.79	6.95
Roadway condition	Dry		Base level			Base level	

	Wet	---	---	---	---	---	---
	Snow/slush	6.07	-8.97	-14.64	7.93	-1.49	-8.91
	Ice	6.04	-8.94	-14.47	5.80	-1.07	-6.55
Crash characteristics							
Manner of collision	Rear-end		Base level			Base level	
	Head-on	-16.54	22.34	47.36	-49.29	-9.44	74.56
	Sideswipe (same/opposite direction)	20.62	-31.51	-46.00	---	---	---
	Angle	---	---	---	-37.86	4.45	45.30
Collision type	Truck with truck		Base level			Base level	
	Passenger car with truck	---	---	---	---	---	---
	Passenger car with passenger car	-2.84	4.15	7.03	-3.44	0.56	3.96

Notes: “NA” represents “not applicable”; “---” represents the variable is not statistically significant at 5% level of significance.

2.6 SUMMARY AND CONCLUSIONS

Traffic safety is an important issue with serious social and financial implications including injuries, fatalities and economic losses. Reducing the number of crashes and their consequences (especially the severe ones) is an important priority for transportation safety professionals. To this end, it is necessary to explore the potential causes of crash severity, so that effective countermeasures can be implemented to alleviate the crash risk.

Crash severity including injury severity and vehicle damage has been widely studied in the literature. Numerous statistical methodologies have been implemented to identify the relationships between different explanatory variables and crash severity. Irrespective of the different model assumptions and structures, failing to capture the dependencies between injury severity and vehicle damage caused by common observed and unobserved factors may lead to the biased coefficient estimates. To address this issue, a copula based ordered probit-ordered probit model is used in this study to jointly model injury severity and vehicle damage by accommodating their dependencies. Furthermore, a parameterized copula structure is used to investigate the varied dependencies between injury severity and vehicle damage across crashes, and the elasticity effects for all independent variables were calculated to explore their effects on the probability of each injury severity and vehicle damage category.

Six copula based models including Gaussian, FGM, Frank, Clayton, Joe and Gumbel copula models and an independent model were tested in this study. The comparison of the model estimations shows that the copula based models had a better goodness-of-fit than the independent model which indicates the existence of dependencies between injury severity and vehicle damage.

Among the copula based models, the Gaussian copula model had the best model performance with the lowest BIC value.

The Gaussian copula model reveals that human factors have significant influences on injury severity. Young drivers are less likely to be associated with severe injuries than others. Males have a lower probability of suffering severe injury severity compared with females. Using alcohol or drug dramatically increases the injuries and using safety restraints considerably decreases the probability of severe injuries. The crash severity on interstate highways is increased due to the higher speed. Four-way stop controlled intersections may be safer than others as both injury severity and vehicle damage are decreased. When compared with normal roadway conditions, adverse surface decreases the crash severity due to the reduced traveling speed. Night time seems to increase the probability of severe vehicle damage but it is not statistically significant for the injury severity model. Compared with the rear-end crashes, head-on crashes increase the probability of severe injuries and both head-on and angle crashes increase the probability of severe vehicle damage. The crash severity for crashes between two passenger cars may be increased due to the larger speed differentials between two vehicles.

The estimated copula parameters offer additional insight about different patterns of dependencies between injury severity and vehicle damage across crashes. The results indicate that dependencies between injury severity and vehicle damage are positive for head-on, angle and sideswipe crashes, while the dependencies are negative for the crashes between two passenger cars. These conclusions indicate that the dependencies between injury severity and vehicle damage can vary across different crashes. In summary, this study offers a more accurate model

structure of predicting crash severity, and it is anticipated that this study can shed light on help develop cost-effective countermeasures to improve traffic safety.

One limitation of the study is that it employs only two vehicle crashes for the analysis. The findings are not directly transferable to crashes involving single vehicles or more than two vehicles. These are avenues for future research. From a practice perspective, the availability of vehicle damage information for roadway crashes might also influence applicability of the proposed framework. However, it is important to recognize that while vehicle damage component of the model might not be employed, the model results obtained for severity analysis can be directly employed. The injury severity estimates obtained through our two dependent variable analysis have been "purified" by considering dependency between the two variables. Hence, the states with no vehicle damage would continue using the injury severity model independently. However, from our analysis, it is evident that considering vehicle damage – an objective indicator of crash severity – might enhance crash severity analysis (35). Therefore, to accurately identify the severity of a crash, compiling vehicle damage is a recommendation from our analysis.

2.7 ACKNOWLEDGMENTS

The authors are grateful to the University of Wisconsin-Madison Traffic Operations and Safety (TOPS) Laboratory for providing the data.

2.8 REFERENCES

1. U.S. Department of Transportation. Federal Highway Administration (FHWA). http://www.fhwa.dot.gov/

2. American Association of State Highway and Transportation Official (AASHTO). http://www.transportation.org/Pages/Default.aspx
3. FHWA-SA-10-005. Federal Highway Administration (FHWA). The U.S. Department of Transportation. 2009. http://safety.fhwa.dot.gov/intersection/resources/fhwasa10005/docs/brief_2.pdf
4. Unites States. NHTSA National Center for Statistic and Analysis. Passenger Vehicle Occupant Fatalities: The Decline for Six Years in a Row from 2005 to 2011. http://www-nrd.nhtsa.dot.gov/Pubs/812034.pdf. 2014.
5. Kim K, P. Pant and E. Yamashita. Measuring Influence of Accessibility on Accident Severity with Structural Equation Modeling. In Transportation Research Record 2236, TRB, National Research Council, Washington, D.C., pp. 1-10. 2011.
6. Wang K. and X. Qin. Using structural equation modeling to measure single-vehicle crash severity. Transportation Research Record. Report No. 14-0801. TRB, National Research Council, Washington, D.C., 2014.
7. Qin X, K. Wang and C. Cutler. Modeling Large Truck Safety Using Logistic Regression Models. Accepted by Transportation Research Record, TRB, National Research Council, Washington, D.C., Paper No. 13-2067. 2013.
8. Dissanayake S.. Comparison of Severity Affecting Factors Between Young and Older Drivers Involved in Single Vehicle Crashes. International Association of Traffic and Safety Sciences, Vol. 28, pp. 48-54. 2004.
9. Ye, F., and D. Lord. Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models Multinomial Logit, Ordered Probit, and

Mixed Logit. In Transportation Research Record 2241, TRB, National Research Council, Washington, D.C., pp. 51-58. 2011.

10. Ghulam H, J. Bhanu and M. Uday. Multinomial Logistic Regression Model for Single-Vehicle and Multivehicle Collisions on Urban U.S. Highways in Arkansas. Journal of Transportation Engineering. Vol. 138. No. 6, pp. 786-797. 2012.
11. Zajac, S. S. and J. N. Ivan. Factors Influencing Injury Severity of Motor Vehicle–Crossing Pedestrian Crashes in Rural Connecticut. Accident Analysis and Prevention. Vol. 35, No. 3, pp. 369–379. 2003.
12. Khattak, A. J., P. Kantor and F. M. Council. Role of Adverse Weather in Key Crash Types on Limited-Access Roadways: Implications for Advanced Weather Systems. In Transportation Research Record 1621, TRB, National Research Council, Washington, D.C., pp. 10–19. 1998.
13. Kockelman, K. M. and Y. J. Kweon. Driver Injury Severity: An Application of Ordered Probit Models. Accident Analysis and Prevention, Vol. 34, No. 3, pp. 313–321. 2002.
14. Abdel-Aty, M. and J. Keller. Exploring the Overall and Specific Crash Severity Levels at Signalized Intersections. Accident Analysis and Prevention, Vol. 37, pp. 417–425. 2005.
15. Christoforou Z., S. Cohen and G. Karlaftis. Vehicle occupant injury severity on highways: An empirical investigation. Accident Analysis and Prevention. Vol. 42, No. 6, pp. 1606-1620. 2010
16. O'Donnell, C. J. and D. H. Connor. Predicting the Severity of Motor Vehicle Accident Injuries Using Models of Ordered Multiple Choice. Accident Analysis and Prevention. Vol. 28, No. 6, pp. 739–753. 1996.

17. Eluru, N., C. R. Bhat, and D. A. Hensher. A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes. Accident Analysis & Prevention. Vol. 40, No.3, pp. 1033-1054. 2008
18. Yasmin. S., N. Eluru, C. R. Bhat and R. Tay. A Latent Segmentation Generalized Ordered Logit Model to Examine Factors Influencing Driver Injury Severity. Analytic Methods in Accident Research 1. pp. 23-38. 2014.
19. Eluru N. Evaluating Alternate Discrete Choice Frameworks for Modeling Ordinal Discrete Variables. Accident Analysis & Prevention. Vol. 55, No. 1, pp. 1-11. 2013.
20. Yasmin. S., and N. Eluru. Evaluating Alternate Discrete Outcome Frameworks for Modeling Crash Injury Severity. Accident Analysis & Prevention. Vol. 59, No. 1, pp. 506-521. 2013.
21. Mooradian, J., J. N. Ivan, N. Ravishanker, and S. Hu. Analysis of Driver and Passenger Crash Injury Severity Using Partial Proportional Odds Models. Accident Analysis and Prevention. Vo. 58, pp. 53-58. 2013.
22. Malyshkina N. and F. Mannering. Markov switching multinomial logit model: An application to accident-injury severities. Accident Analysis and Prevention. Vol. 41. No. 4, pp. 829-838. 2009.
23. Chen, F., and S. Chen. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis and Prevention. Vol. 43, No. 5. pp. 1677-1688. 2011.
24. Moore D. N., W. Schneider, P. T. Savolainen and M. Farzaneh. Mixed logit analysis of bicycle injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis and Prevention. Vol. 43. No. 3. pp. 621-630. 2011.

25. Milton J. C., V. N. Shankar and F. L. Mannering. Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident Analysis and Prevention. Vol. 40. No. 1. pp. 260-266. 2008.
26. Kim J. K., G. Ulfarsson, V. Shankar and F. Mannering. A note on modeling pedestrian injury severity in motor vehicle crashes with the mixed logit model. Accident Analysis and Prevention. Vol. 40. No. 5. pp. 1695-1702. 2010.
27. Abay, K.A., R. Paleti, and C. R. Bhat. The Joint Analysis of Injury Severity of Drivers in Two-Vehicle Crashes Accommodating Seat Belt Use Endogeneity. Transportation Research Part B, Vol. 50, pp. 74-89. 2013.
28. Savolainen, P. T., F. L. Mannering, D. Lord and M. A. Quddus. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. Accident Analysis and Prevention. Vol. 43, No. 5, pp. 1666–1676. 2011.
29. Khattak A., R. Schneider and F. Targa. Risk Factors in Large Truck Rollovers and Injury Severity: Analysis of Single-vehicle Collisions. CD-ROM. Transportation Research Board of the National Academics, Washington, D.C., Paper No 03-2331. 2003.
30. Schorr J., S. Hamdar and T. Vassallo. Collision Propensity Index for Un-signalized Intersections: A structural Equation Modeling Approach. CD-ROM. Transportation Research Board of The National Academics, Washington, D.C., Paper No. 13-3915. 2013.
31. Hassan H. and M. Abdel-Aty. Exploring the safety implications of young drivers' behavior, attitudes and perceptions. Accident Analysis and Prevention. Vol. 50, pp. 361-370. 2012.
32. Hamdar S., H. Maahmassani and R. Chen. Aggressiveness propensity index for driving behavior at signalized intersections. Accident Analysis and Prevention. Vol. 40, pp 315-326. 2008.

33. Ambak K., R Ismail, R. Abdullah and M. Borhan. Prediction of Helmet Use among Malaysian Motorcyclist Using Structural Equation Modeling. Australian Journal of Basic and Applied Sciences. Vol. 4, No. 10, pp. 5263-5270, 2010.
34. Lee, J., J. Chung and B. Son. Analysis of Traffic Accident Size for Korean Highway Using Structural Equation Models. Accident Analysis and Prevention, Vol. 40, pp. 1955-1963. 2008.
35. Qin X, K. Wang, and C. Cutler. Analyzing Crash Severity Based on Vehicle Damage and Occupant Injuries. In Transportation Research Record 2386, National Research Council, Washington, D.C., pp. 95-102. 2013.
36. Huang, H., H. C. Chin and M. M. Haque. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention. Vol. 40, pp. 45-54. 2008.
37. Quddus, M., R. B. Noland and H. C. Chin. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Journal of Safety Research, 33, pp. 445-462. 2002.
38. Washington, S., M. Karlaftis, F. L. Mannering. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL. 2011.
39. Yasmin S., N. Eluru, A. Pinjari and R. Tay. Examining Driver Injury Severity in Two Vehicle Crashes-A Copula Based Approach. Accident Analysis and Prevention. Vol. 66. pp. 120-135. 2014.
40. Pourabdollahi Z., B. Karimi and A. Mohammadian. Joint Model of Freight Mode and Shipment Size Choice. In Transportation Research Record 2378, TRB, National Research Council, Washington, D.C., pp. 84–91. 2013.

41. Sener I., N. Eluru and C. R. Bhat. On Jointly Analyzing the Physical Activity Participation Levels of Individuals in A Family Unit Using a Multivariate Copula Framework. Journal of Choice Modelling. Vol. 3, No. 3, pp. 1-38. 2010.
42. Eluru N., R. Paleti, R. Pendyala and C. Bhat. Modeling Multiple Vehicle Occupant Injury Severity: A Copula-Based Multivariate Approach. In Transportation Research Record 2165, TRB, National Research Council, Washington, D.C., pp. 1–11. 2010.
43. Rana T., S. Sikder and A. Pinjari. A Copula-Based Method to Address Endogeneity in Traffic Crash Injury Severity Models: Application to Two-Vehicle Crashes. In Transportation Research Record 2147, TRB, National Research Council, Washington, D.C., pp. 75–87. 2010.
44. Bhat C. R., and N. Eluru. A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling. Transportation Research Part B: Methodological. Vol. 43. No. 7. pp. 749-765. 2009.
45. Trivedi P., D. Zimmer. Copula Modeling: An Introduction for Practitioners. Foundations and Trends in Econometrics, Vol. 1, No. 1, pp. 1-110. 2007.
46. Model Minimum Uniform Crash Criteria Fourth Edition. U.S. Department of Transportation, Washington, D.C. 2012. http://www.mmucc.us.
47. Law Enforcement Officer's Instruction Manual for Completing the Wisconsin Motor Vehicle Accident Report Form (MV 4000); WisDOT, Division of Motor Vehicles, 1998.
48. Eluru, N. and C. R. Bhat. A Joint Econometric Analysis of Seat Belt Use and Crash-Related Injury Severity. Accident Analysis and Prevention. Vol. 39, No. 5, pp. 1037-1049. 2007.

2.9 APPENDIX: PARAMETER ESTIAMTION AND EXPLANATION

In the copula based model, the probability of each injury severity level is shown in Equation (2-2) as:

$$Pr(y_q = j) = \phi(\tau_j - \alpha' x_q) - \phi(\tau_{j-1} - \alpha' x_q)$$

the probability of each vehicle damage level is shown in Equation (2-4) as:

$$Pr(u_q = k) = \Lambda(\psi_k - \beta' z_q) - \Lambda(\psi_{k-1} - \beta' z_q)$$

the probability of each combination of injury severity and vehicle damage level is shown in Equation (2-6) as:

$$Pr(y_{qj} = j, u_{qk} = k)$$

$$= C_{\theta q}(U_{qj}, U_{qk}) - C_{\theta q}(U_{qj}, U_{qk-1}) - C_{\theta q}(U_{qj-1}, U_{qk}) + C_{\theta q}(U_{qj-1}, U_{qk-1})$$

and the dependency between injury severity and vehicle damage is shown in Equation (2-7) as:

$$\theta_q = f_n(\gamma' s_q)$$

Focusing on the three equations above, the unknown parameters to be estimated are: α' and τ_j in the injury severity component, β' and ψ_k in vehicle damage component, and finally γ' in the dependency component. α' represents the estimated coefficients for independent variables in the injury severity component, a positive value of α' represents a propensity to increase the injury and vice-versa for a negative value of α'. τ_j represents the threshold of the ordered Probit model to determine the estimated injury severity level for a specific crash. β' represents the estimated coefficients for independent variables in the vehicle damage component, a positive value of β' represents a propensity to increase the vehicle damage and vice-versa for a negative value of β'. ψ_k represents the threshold of the ordered Probit model to determine the estimated vehicle damage level for a specific crash. γ' is the estimated coefficients for the dependency between injury severity and vehicle damage caused by the common unobserved factors. Specifically, a

positive γ' indicates that when injury severity increases, the vehicle damage increases too for the specific type of crashes, and a negative γ' indicates that when injury severity increases, the vehicle damage decreases for the specific type of crashes.

Finally, the likelihood function of joint probability of injury severity and vehicle damage is shown in Equation (2-8) as:

$$L = \prod_{q=1}^{Q}\left[\prod_{j=1}^{J}\prod_{k=1}^{K}\{Pr(y_q = j, u_q = k)\}^{\omega_{qkj}}\right]$$

the logarithm of the above likelihood function is coded using the GAUSS programming language, and the maximum likelihood estimation approach is use to estimate the model parameters.

3 PREDICTING LOCAL ROAD CRASHES USING SOCIO-ECONOMIC AND LAND COVER DATA

3.1 INTRODUCTION

A Safety Performance Function (SPF) is an equation used to predict crash counts at a location as a function of exposure and other roadway characteristics (*e.g.* number of lanes, lane width, shoulder width) (1). One of the uses for SPFs is estimating the expected number of crashes on traffic facilities to identify road locations with higher crash potential for safety improvements, select and implement cost-effective countermeasures to reduce future crashes (2). SPFs are often developed for different traffic facilities such as road segments and intersections. Local roads owned and operated by local entities including towns, counties and tribal governments play an important role in the roadway network, as approximately 60 percent of all road miles in the U.S. are maintained by these jurisdictions (3). A recent Iowa study (4) reported that local roads had higher crash rates compared to primary roads under State jurisdiction and the reported local road crash rate was 1.5 times higher than that of primary roads from 1974 to 2000. As a result, traffic safety on local roads is important to both traffic safety organizations and engineers. Given this situation, it is important to develop accurate tools to predict the number of crashes occurred on local roads to support identifying sites with promise for safety improvements and selecting and implementing effective countermeasures to reduce future crash volume or severity.

The Highway Safety Manual (HSM) (1) provides SPFs for two lane rural highways, multilane rural highways, urban and suburban arterials, freeways and freeway ramp junctions. The SPFs in HSM were estimated using data collected from a limited number of States in the USA, including Washington, California, Minnesota, Texas, Michigan, North Carolina and Illinois. Because crash

relationships in these states are not necessarily representative of those in the entire country, the HSM recommends a calibration procedure to adjust the predicted crash counts for individual jurisdiction in using the prediction from the SPF. The HSM SPFs include traffic counts for intersections or roadway segments as the most critical variables in accurately predicting the number of crashes (1, 5, 6). This presents a problem for roads under local jurisdiction, where traffic counts are generally not available because it is economically impractical to implement traffic counting programs for so many facilities on which the traffic volume is typically below 400 per day (4). As well, the data sets used to estimate the two-lane road models in the HSM do not include roads with traffic volumes as low as are usually found on many city or town jurisdiction streets and roads. In order to implement highway safety improvement strategies on these low volume local roads, new crash prediction approaches are desirable, in which the traffic counts are not required.

The objective of this study was to estimate SPFs for both intersections and segments on roads under local jurisdiction in the State of Connecticut using demographic data as a surrogate for traffic count data. The SPFs are estimated at the level of Traffic Analysis Zone (TAZ), instead of the intersection or roadway segment level. The intersection counts (*i.e.* the number of city/town road intersections in a TAZ) and segment mileage (*i.e.* total city/town roadway length in a TAZ) are used as exposure in this study in lieu of traffic volume. Demographic records such as population, total retail and non-retail employment, household income and vehicle availability work in tandem with the exposure to predict the estimated crash counts. To account for data and crash relationship heterogeneity, the TAZs in the entire state are categorized into six clusters based on the percentage of three land cover categories – high, medium and low intensities – and

the population density (*i.e.* the number of population per km^2). A different SPF was estimated for each cluster, and the similarities and differences among these functions are discussed. We also discuss how to apply the functions as a network screening tool.

It is noted that the term "local" can mean different things depending on the context in which it is used. In one context it can refer to one level of the hierarchical functional classification scheme (arterial, collector and local). It can also be used to refer to the level of agency jurisdiction responsible for a road facility (state, county, local). It is possible for the same road to be called two different things, for example a "collector" in the first context, but "local" in the second. To avoid this confusion, we use the word "local" for the first context and "city or town" in the second context. Note also that there are no roads in the State of Connecticut under county jurisdiction.

3.2 LITERATURE REVIEW

SPFs have been estimated for city or town roads by various researchers at two levels: the facility level (*e.g.* roadway segment and intersection) and the zonal level (*e.g.* TAZ). Among facility level models, Vogt (6) provides a good review of the factors associated with crashes on city or town roads according to past research studies. These include channelization (right and left turn lane), number of driveways, sight distance, intersection angle, median width, surface width, shoulder width, signal characteristics, lighting, roadside condition, truck percentage in the traffic volume, posted speed, and weather. Most research on two-lane roads confirms traffic volume as the major explanatory factor for traffic crashes, which is unfortunate for the cases where the traffic volume is not available (7, 8). There is little literature on investigating alternative exposure measures in addition to or in place of traffic volume for predicting crashes. Bindra *et al.*

(9) considered the use of geographic information system (GIS) land use inventories to supplement traffic volumes as exposure for estimating SPFs for predicting segment-intersections crashes for rural two-lane and urban two-and four-lane undivided roads. They concluded that the number of trips generated and the land use data (*i.e.*, population, retail and non-retail employment, and driveway data) were good predictors for estimating segment-intersection crashes, that is, crashes on segments located at minor roads and driveways without traffic counts.

Zonal SPFs (ZSPFs), of which the most popular is TAZ level, make use of highly available zonal-level variables (10). Among the studies focusing on developing TAZ-level SPFs, Pulugurtha *et al.* (11) used socioeconomic and network variables to develop TAZ level SPFs to estimate the crash counts by severity level (injury and property damage only crashes). Ladron de Guevara *et al.* (12), Lovegrove and Sayed (13), Lovegrove (14) and Hadayeghi *et al.* (15) developed TAZ level SPFs to estimate the number of both intersection and segment crashes. Factors such as population density, the number of employees and the intersection density were considered as predictors for the number of crashes. Furthermore, Khondakar *et al*. (16) found that TAZ level SPFs can safely be transferred both temporally and spatially. Noland and Quddus (17) showed that TAZs with high employment density had more traffic crashes, whereas in urbanized areas with more densely populated TAZs fewer crashes were observed. Jin *et al.* (18) identified that besides traditional variables such as segment length, structure of roadway network should be considered in developing TAZ-level SPFs to improve prediction accuracy. Several studies developed TAZ-level SPFs using number of trips generated inside of each TAZ. Naderan and Shahi (19), Abdel-Aty *et al.* (20) found that number of trips generated have significant impacts on TAZ-level crashes.

Recently, an analysis tool (PLANSAFE) was developed on a National Cooperative Highway Research Program (NCHRP) project (21) to predict the expected crash counts by TAZ. The predictors include population, employment and some land use intensity variables. The purpose was to use the predicted crash counts as one of the measures of effectiveness to select the most cost-effective transportation improvement plan. Another study of TAZ level SPFs by Pirdavani *et al.* (10) considered establishing an association between observed crashes and a set of predictor variables in each TAZ. The study compared models using two different exposures - VHT (total daily vehicle hours traveled) and VKT (total daily vehicle kilometers traveled) along with network and socio-demographic variables. The results show that the model containing the combination of two exposures outperformed the models containing only one of the exposure variables. Lee *et al.* (22) applied a multivariate Poisson Lognormal crash modeling to simultaneously estimate motor vehicle crashes, bicycle crashes and pedestrian crashes by using several socio-demographic variables in each TAZ. The study illustrates that the number of households, employments and hotels etc. are positively associated with three types of crash counts. Except for TAZ-level SPFs, some studies have investigated SPFs on other macroscopic levels, such as block group (23, 24), state level (25), grid structure level (26) and county level (27, 28, 29).

These zonal level SPFs are all able to predict expected crash frequencies without traffic volume, however most of them estimate the number of crashes using network and social-demographic variables, *etc*., without accounting for the data and crash heterogeneity among different types of TAZs or zones. To address this issue, our study focuses on estimating TAZ level SPFs that do

not require ADT counts for city and town jurisdiction roads by different categories of TAZ. The TAZs were clustered into different categories using a data mining technology (K-means clustering analysis), based on their land-use intensities and population density. Socio-demographic data and roadway network data such as population, employment, income, car ownership, number of city/town jurisdiction road intersections and total city/town road length inside the TAZ are used to predict crash counts. The intention is for some of the variables to serve as surrogates for actual traffic counts which are generally not available for these roads.

The remainder of the paper is organized as follows. The next section presents the methodology and the process of data collection. The third section describes the estimation of SPFs and the results. The final section discusses how to use the estimated SPFs as a network screening tool.

3.3 METHODOLOGY AND DATA PREPARATION

Our procedure for the estimation of TAZ level SPFs for city and town roads requires four types of data at the TAZ level: roadway network shape features, demographic records, geographic/land cover features and crash records. We chose to use the TAZ structure defined by CT DOT for statewide planning purposes to take advantage of the extensive array of demographic data available by TAZ. Below is a brief description of the required data and data sources.

3.3.1 ROADWAY NETWORK SHAPE FEATURES

The number of intersections and the total length of roadways under city or town jurisdiction were extracted from the 2010 Census TIGER/LINE files for Connecticut (30). The original TIGER/LINE files contained correction of errors, such as typos for roadway name and discrepancies in the network representation of some road links. The number of intersections and the total length of roadways under city or town jurisdiction were calculated for each TAZ.

Details about our procedures for calculating the number of intersections and the total length of roadways are provided in the Appendix to the project final report (31).

3.3.2 TAZ LEVEL DEMOGRAPHIC RECORDS

TAZ level demographic records were collected from the Census Transportation Planning Package Database (32). They include population, retail and non-retail employment, households, vehicles and average household income summarized by TAZ and used as the independent variables in safety performance functions. In the 2010 census, 1806 TAZs were defined for the State of Connecticut. Two of these TAZs were apparently defined to represent special generators and have no population or employment, so they were eliminated from the analysis. The remaining 1804 TAZs were used to estimate the SPFs.

3.3.3 TAZ LEVEL GEOGRAPHIC/LAND COVER FEATURES

Land-cover information was collected from the 2011 NLCD (National Land Cover Database) (33). We calculated the proportion of land area in three developed land-use categories – low, medium and high intensity development – as defined by USGS (33). All developed areas contain a mixture of vegetation and impervious surfaces (*e.g.*, buildings, roadways), where development intensity reflects differences in the relative proportions of these cover types. The classification system employed by the 2011 NLCD defines low intensity areas as having 20%-49% impervious cover, medium intensity areas as having 50%-79% impervious cover, and high intensity areas as having greater than 80% impervious cover (33). These values along with the population density were used to categorize the TAZs into homogeneous groups using K-means clustering analysis (discussed in the next section). Originally we used only the land cover intensities, but we found that adding the population density helped to correct aberrant cluster assignments for unique development sites (*e.g.*, airports).

3.3.4 CRASH RECORDS AND INTEGRATION OF CRASH TO TAZ

Intersection and segment crash records were collected from the Connecticut Crash Data Repository (34). We gathered counts of K (fatal injury), A (incapacitating injury) and B (non-incapacitating injury) intersection and segment crashes occurring on roads under city and town jurisdiction in Connecticut from 2010 to 2012. Crashes at intersections with one or more approaches maintained by the State were not included. As requested by the Technical Advisory Committee for the project, we excluded property damage only PDO (O) and minor injury (C) crashes because they lead to less serious consequences and are also subject to underreporting (PDO's in particular). Also, cities and towns were not required to report PDO crashes in 2011, so the dataset would have been incomplete if we included them. In total, 5403 intersection crashes and 5502 segment crashes were extracted.

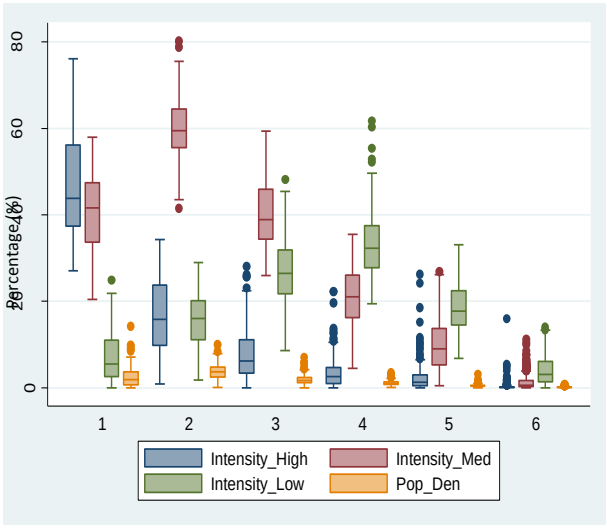
Intersection and segment crashes were assigned to TAZs based on their locations. If the crash was located on the boundary of more than one TAZ, it was evenly assigned between the two TAZs on both sides of the road where the crash occurred in the case of a segment crash. For an intersection crash on the intersection of several TAZs, it was equally assigned among all TAZs that touch the intersection (an intersection crash would be evenly assigned among four TAZs for a four-way intersection that forms the corner of four TAZs). Details about our procedures for assigning crashes are provided in the Appendix to the project final report (31).

3.3.5 CLUSTERING OF TAZs

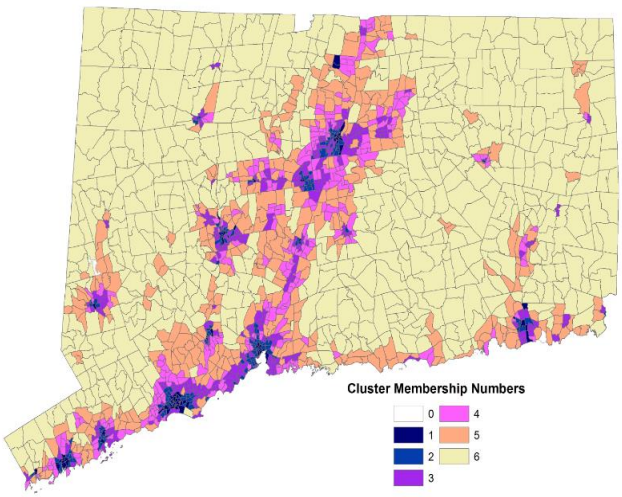
Clustering analysis seeks to maximize the similarity of contents within the same cluster and the dissimilarity of elements between clusters (35). K-means clustering analysis (35, 36, 37) is a traditional distance-based technique which has a limitation that a distance measured objective

function is required to be pre-determined. The second issue of this methodology is that it requires large memory demands especially for a large dataset (35). To account for these issues, the latent class clustering (LCC) analysis or finite mixture model (FMM) was applied by numerous studies, as it doesn't require selecting a distance measure. However, LCC is a model-based technology which is not appropriate for our data, as there is no dependent variable in our clustering process. Therefore, considering the simplicity and data structure, K-means clustering analysis with the Euclidean distance measured objective function (38) was selected to categorize the TAZs into homogeneous groups using the three land cover intensities and the population density. Different numbers of clusters were respectively tested, and the Calinski and Harabase pseudo-F index (39) was used to select the final number of clusters. The larger the Calinski and Harabase pseudo-F index, the more accurate is the clustering analysis.

The optimum number of clusters was found to be six. Figure 3.1(a) shows the distributions of the three land-use intensities and the population density among the six clusters. The overall land-use intensity and the population density decrease from cluster 1 to cluster 6. The number of TAZs assigned into cluster 1 through cluster 6 is 80, 161, 270, 284, 382 and 627, respectively. Figure 3.1(b) shows the distribution of the six clusters across the state. Note that two TAZs with legend 0 in the western and southeastern areas were eliminated in estimating the safety performance functions, as these two TAZs have no population. Cluster 1 has the lowest number of TAZs, and is the most urbanized in nature, and cluster 6 is the most common cluster type and is the most rural in nature. Clusters 2 through 5 represent areas with decreasing levels of urbanization. The areas with higher land-use intensities (those with the darkest shading and colors on the map) are mainly located in the central and southern parts of the state.

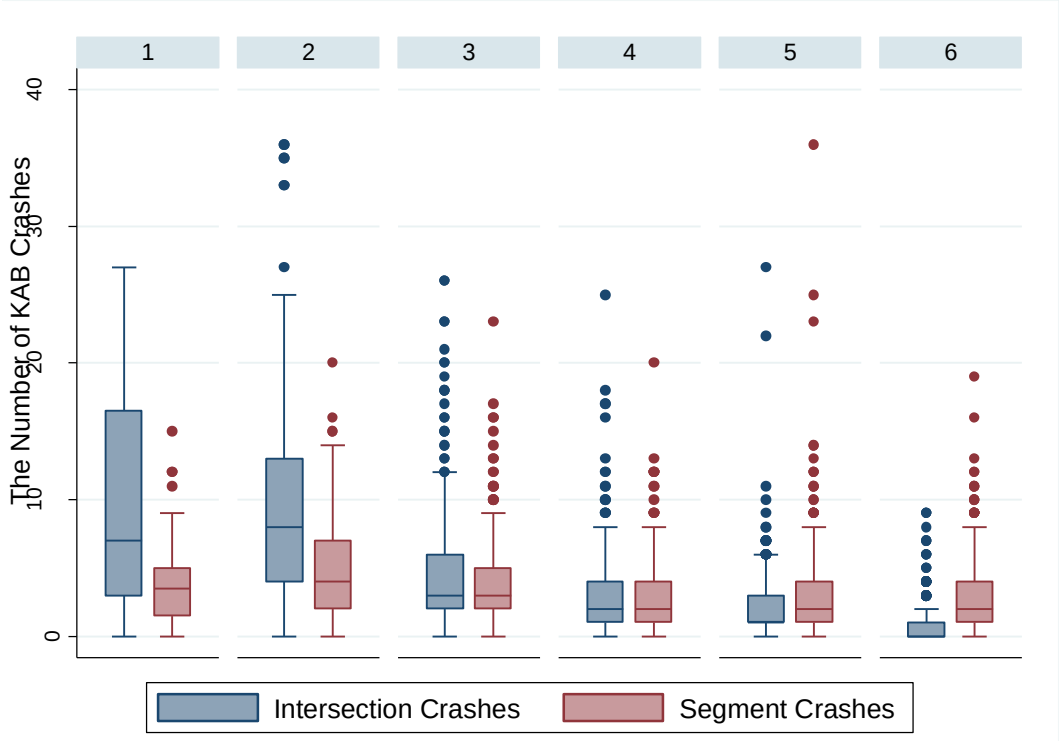


(a) land-use intensities and population density distributions by cluster



(b) cluster distribution over Connecticut

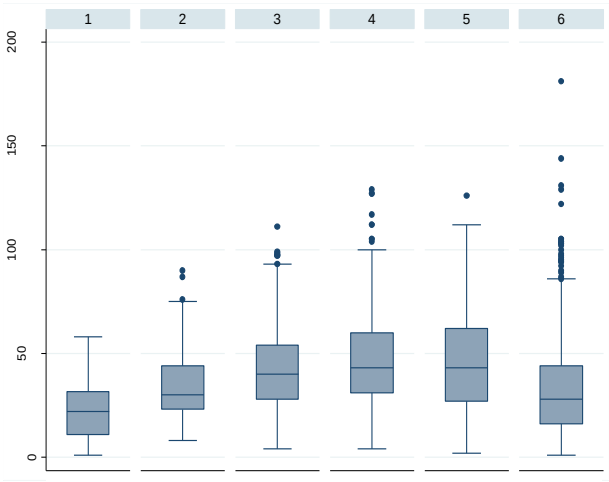
FIGURE 3.1 Clustering Results and Cluster Distribution



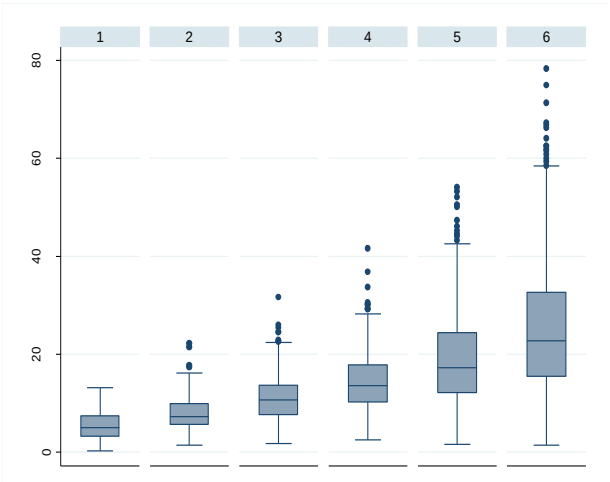
(*the boxplot from left to right under each cluster is related to intersection crashes and segment crashes*)

FIGURE 3.2 Distributions of KAB Crashes by Cluster

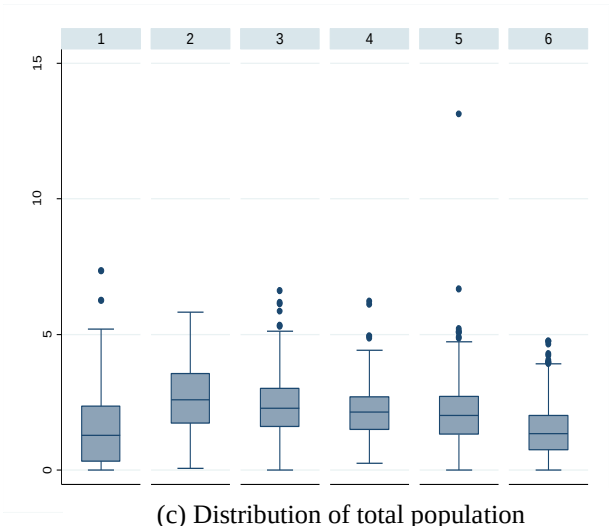
Figure 3.2 illustrates the distribution of KAB crashes by cluster. Comparing the two types of crashes, there are substantially more intersection crashes than segment crashes in clusters 1, 2 and 3, but fewer intersection crashes than segment crashes in clusters 5 and 6. The two types of crashes have nearly the same distributions in cluster 4. Figure 3 and Figure 4 display the distributions of the number of intersections, city or town roadway mileage and demographic variables by cluster. The number of intersections increases from cluster 1 to cluster 5, and then decreases to cluster 6. The roadway mileage increases consistently from cluster 1 to cluster 6. The average household income slightly increases from cluster 1 to cluster 6. Cluster 1 has the highest average numbers for both retail and non-retail employment, and cluster 6 has the lowest numbers. One important finding is that the distribution patterns are similar among population (Figure 3.3(c)), households (Figure 3.3(d)) and vehicles (Figure 3.4(a)). This is caused by the high correlation among these three factors, which was also verified by a correlation test. The selection and application of these three correlated variables is discussed under SPF development.



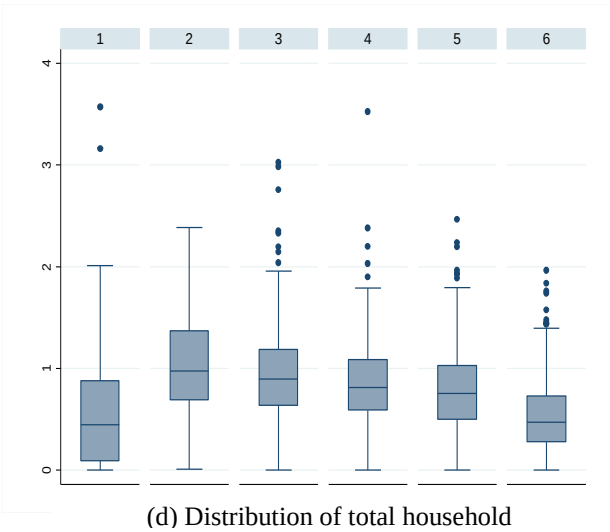
(a) Distribution of the number of intersections



(b) Distribution of city or town roadway mileage

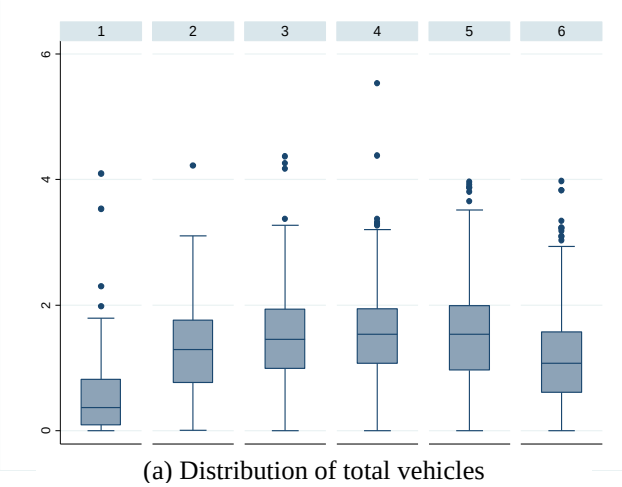


(c) Distribution of total population

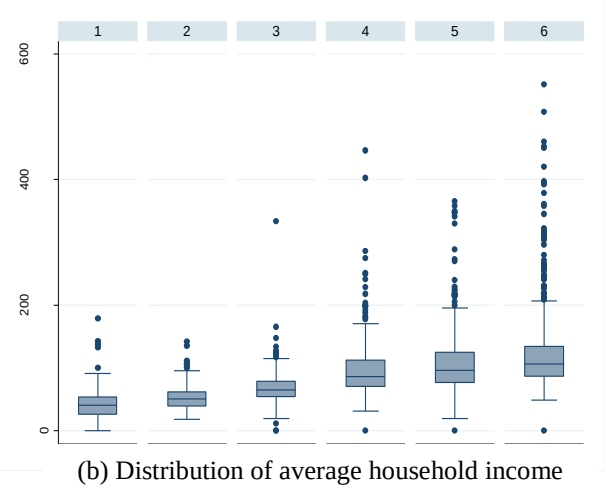


(d) Distribution of total household

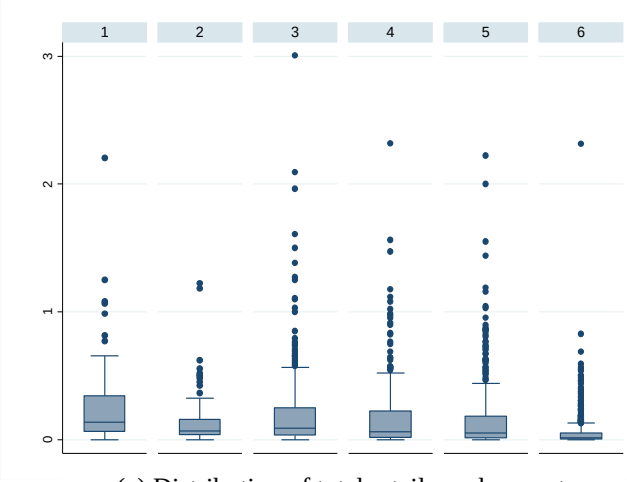
FIGURE 3.3 Distributions of Independent Variables by Cluster



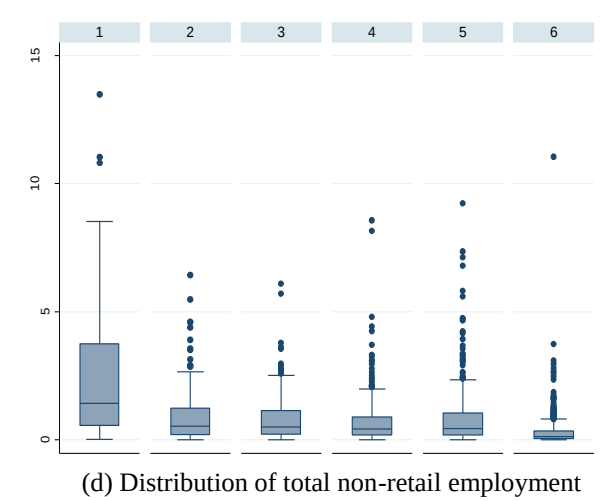
(a) Distribution of total vehicles



(b) Distribution of average household income



(c) Distribution of total retail employment



(d) Distribution of total non-retail employment

FIGURE 3.4 Distributions of Independent Variables by Cluster (Continued)

3.3.6 STATISTICAL METHODOLOGY

Safety performance functions were estimated to predict the number of city and town road intersection and segment crashes in each TAZ. The number of crashes is estimated by count regression models, such as the Poisson regression model, formulated as (40):

$$Prob[y_i|\mu_i] = \frac{exp(-\mu_i)\mu_i{}^{y_i}}{y_i!} \quad (3\text{-}1)$$

where $Prob[y_i|\mu_i]$ is the probability of y crashes occurring at TAZ i and μ_i is the expected number of crashes at TAZ i. Given a vector of covariates X_i, which describes the demographic

and roadway characteristics of a TAZ i, and a vector of estimable coefficients β, the μ_i can be estimated by the equation:

$$ln(\mu_i) = \boldsymbol{\beta X_i} \tag{3-2}$$

The limitation of the Poisson model is that the variance of the data is constrained to be equal to the mean, *i.e.*:

$$Var(y_i) = E(y_i) = \mu_i \tag{3-3}$$

This constraint might be questionable as the variance of crash data is usually greater than the mean, which is known as over-dispersion (40). The negative binomial regression model addresses this issue, which is derived by rewriting Equation (3-2) such that:

$$\mu_i = exp(\boldsymbol{\beta X_i} + \varepsilon_i) \tag{3-4}$$

where $exp(\varepsilon_i)$ is an error term assumed to follow a gamma distribution with mean 1 and variance σ^2. The distribution of the negative binomial model has the form (40):

$$Prob[y_i|\mu_i] = \frac{\Gamma\left[\left(\frac{1}{\sigma}\right)+y_i\right]}{\Gamma\left(\frac{1}{\sigma}\right)y_i!}\left[\frac{\frac{1}{\sigma}}{\left(\frac{1}{\sigma}\right)+\mu_i}\right]^{\frac{1}{\sigma}}\left[\frac{\mu_i}{\left(\frac{1}{\sigma}\right)+\mu_i}\right]^{\mu_i} \tag{3-5}$$

where Γ is a gamma function; the variance of the negative binomial model can be written as follows:

$$Var(y_i) = \mu_i(1 + \sigma\mu_i) = \mu_i + \sigma{\mu_i}^2 \tag{3-6}$$

We define the function for the predicted intersection crashes at TAZ i as follows:

$$\mu_{int,i} = Y{I_i}^{\beta_I} exp(\beta_0 + \beta_P P_i + \beta_R R_i + \beta_N N_i + \beta_V V_i + \beta_C C_i + \beta_H H_i) \tag{3-7}$$

Where

$\mu_{int,i}$ = predicted intersection crashes in TAZ i

Y = the number of years in the time period

I_i = the number of intersections in TAZ i

P_i = the population of TAZ *i*

R_i = the total retail employment of TAZ *i*

N_i = the total non-retail employment of TAZ *i*

V_i = the number of vehicles in TAZ *i*

C_i = the average income in TAZ *i*

H_i = the number of households in TAZ *i*

βs = the estimated parameters

We define the function for the predicted segment crashes at TAZ *i* as follows:

$$\mu_{seg,i} = YL_i{}^{\beta_L} exp(\beta_0 + \beta_P P_i + \beta_R R_i + \beta_N N_i + \beta_V V_i + \beta_C C_i + \beta_H H_i) \quad (3\text{-}8)$$

Where

$\mu_{seg,i}$ = predicted segment crashes in TAZ *i*

L_i = the mileage of roadways under local jurisdiction in TAZ *i*

and the remaining variables are as defined above.

3.4 VARIABLE SELECTION AND SPF RESULTS

The SPFs were estimated at the TAZ level for each cluster type. One statewide SPF using the aggregate data (*i.e.,* for all TAZ's without splitting by cluster) was also estimated for comparison purposes. When estimating each function, the observations by TAZ were randomly divided into two parts: one part including ninety percent of the observations was used to estimate the function; and the other part including the remaining ten percent of the observations was used to evaluate the prediction performance of the function. Three functions, each using one of the correlated independent variables at a time (population, number of households and number of vehicles), were estimated for both intersection and segment crashes. We checked for correlation among the variables included in each model; no significant correlation was found. These three functions

were compared according to the model goodness-of-fit (Akaike Information Criterion-AIC and Bayesian Information Criterion-BIC) to determine which one performed best for each cluster and for the statewide database. The number of crashes was predicted using both estimation and prediction datasets for the entire state using the cluster-based functions and the statewide function to test the efficacy of each approach. Function performance for each cluster and the statewide database was compared using two measures of effectiveness (MOEs), Mean Absolute Deviation (MAD) and Mean Squared Predictor Error (MSPE), proposed by Oh *et al.* (41). These criteria are calculated as:

$$AIC = 2K - 2\,ln(LL) \tag{3-9}$$

$$BIC = K * ln(N) - 2ln(LL) \tag{3-10}$$

$$Mean\ Absolute\ Deviation\ (MAD) = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y_i} - y_i| \tag{3-11}$$

$$Mean\ Squared\ Predictor\ Error\ (MSPE) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2 \tag{3-12}$$

Where

K = the number of estimated parameters

LL = the maximized value of model likelihood function

N = the number of observations

$\widehat{y_i}$ = the predicted number of crashes at TAZ i

y_i = the observed number of crashes at TAZ i

The smaller the AIC, BIC, MAD or MSPE value, the better is the function performance. Table 3.1 shows the goodness-of-fit of the cluster based SPFs and Statewide SPFs including one of the correlated variables at a time. Due to the poorer performance of the function using the number of vehicles, only the functions including population or the number of households are presented here. For the statewide SPF, both intersection and segment SPFs have lower AIC and BIC values

using population than using households. For the intersection SPF, the function for clusters 2, 3 and 4 have a lower AIC or BIC value using population as an independent variable than that using the number of households, while the reverse is observed for clusters 1, 5 and 6. The segment SPFs for all clusters have lower AIC and BIC values using population than using households.

TABLE 3.1 Goodness-of-fit of the Cluster Based and Statewide SPFs

	Intersection SPF				Segment SPF			
Cluster SPF	**Population**		**Households**		**Population**		**Households**	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
1	432	448	428	444	330	346	334	350
2	887	908	896	917	692	713	718	739
3	1,231	1,256	1,246	1,271	1,081	1,105	1,109	1,134
4	1,110	1,135	1,120	1,145	1,051	1,075	1,063	1,088
5	1,220	1,247	1,219	1,246	1,475	1,502	1,489	1,516
6	1,247	1,278	1,246	1,277	2,120	2,151	2,125	2,155
Statewide SPF	6,935	6,972	6,977	7,015	6,826	6,863	6,970	7,008

Table 3.2 displays the SPF prediction performance for the statewide and cluster-based functions using both estimation data and prediction data. The cluster-based SPFs using either population or households outperform the statewide SPF in crash prediction, as they have a lower MAD or MSPE value for both estimation data and prediction data. This is to be expected, as it has the possibility of accounting for heterogeneity related to land cover intensity. Furthermore, the cluster-based SPFs with population slightly outperform the SPFs with the number of households. Additionally, it seems that the SPF performance using the prediction data is even better than that

using the estimation data. This may be due to the smaller size of the prediction data set, but it also demonstrates that there is no over-fitting to the estimation data, and that the functions are transferable within Connecticut. Therefore, considering all of these MOEs (model fit and prediction), the cluster-based SPFs with population were selected.

TABLE 3.2 SPF Prediction Performance

MOEs	Statewide SPF (Population)	Statewide SPF (Households)	Cluster-based SPF (Population)	Cluster-based SPF (Households)
Intersection SPF				
MAD Estimation	2.65	2.72	1.95	1.95
MAD Prediction	2.65	2.74	1.62	1.75
MSPE Estimation	18.25	20.72	11.14	11.29
MSPE Prediction	13.29	14.95	6.41	7.50
Segment SPF				
MAD Estimation	2.00	2.01	1.77	1.87
MAD Prediction	1.52	1.58	1.30	1.47
MSPE Estimation	8.28	9.13	7.55	7.62
MSPE Prediction	4.00	4.48	3.51	3.74

Table 3.3 shows the coefficient estimates for the intersection SPFs using population as a predictor. Coefficients for all other models are omitted here for brevity; they may be found in the Appendix to the Final Report (31). The first row in each table cell is the coefficient, the second row is the p-significance, and coefficients shown in bold are statistically significant with 95% confidence. With respect to the six cluster-based functions, the number of intersections (exposure

surrogate for intersection SPFs) was not statistically significant in the cluster 2, 3 and 4 functions. The effect of total population on number of intersection crashes is shown to be positive in all functions (as expected), except for clusters 5 and 6, in which it was not statistically significant. The amount of retail employment is positively associated with the number of intersection crashes in the functions for cluster 4, 5 and 6. The amount of non-retail employment is positively associated with the number of intersection crashes for cluster 1, 2 and 6. The number of intersection crashes decreases with the increase of average household income in the first five cluster functions, but increases in the cluster 6 function.

TABLE 3.3 Coefficient Estimates for KAB Intersection Crashes

Variables	Coefficient Estimates by Cluster					
	1	**2**	**3**	**4**	**5**	**6**
Intercept	**-1.275**	0.270	-0.150	**-0.984**	**-2.688**	**-4.908**
	(0.001)	*(0.487)*	*(0.717)*	*(0.044)*	*(0.000)*	*(0.000)*
Log (number of intersections)	**0.682**	0.170	0.078	0.040	**0.606**	**0.844**
	(0.000)	*(0.225)*	*(0.587)*	*(0.810)*	*(0.000)*	*(0.000)*
Population (*1000)	**0.161**	**0.282**	**0.360**	**0.372**	0.054	0.129
	(0.014)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.368)*	*(0.145)*
Retail employment (*1000)	0.196	-0.295	-0.221	**0.462**	**0.845**	**0.992**
	(0.530)	*(0.451)*	*(0.261)*	*(0.045)*	*(0.000)*	*(0.000)*
Non-retail employment (*1000)	**0.090**	**0.182**	0.121	-0.003	-0.064	**0.174**
	(0.003)	*(0.000)*	*(0.072)*	*(0.966)*	*(0.195)*	*(0.008)*
Average household income	-0.005	**-0.013**	**-0.010**	-0.002	**-0.003**	**0.002**
(*1000)	*(0.067)*	*(0.000)*	*(0.000)*	*(0.240)*	*(0.009)*	*(0.001)*
Overdispersion	**0.258**	**0.280**	**0.422**	**0.616**	**0.357**	**0.227**
	(0.000)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Deviance/DF	**1.090**	**1.001**	**0.899**	**0.832**	**0.874**	**0.802**

Notes: first row is the coefficient, second row is the p-significance, and bold coefficients are statistically significant at 5% level of significance.

Table 3.4 shows the coefficient estimates for the segment SPFs. Similar to the intersection SPFs, the association between the exposure surrogate, *i.e.* city or town roadway length, and the number

of segment crashes is positive in all six functions, but is only statistically significant in clusters 1, 5 and 6. The coefficient for population is positive and significant in all six cluster-based functions. The retail employment is statistically significant in clusters 3, 4 and 5, and the non-retail employment is statistically significant in clusters 1, 2 and 3. The number of segment crashes decreases with the increase of average household income in the first five cluster functions, but increases in cluster 6 function, which is consistent with the intersection SPFs.

TABLE 3.4 Coefficient Estimates for KAB Segment Crashes

Variables	Coefficient Estimates by Cluster					
	1	**2**	**3**	**4**	**5**	**6**
Intercept	**-3.648**	-1.769	-1.300	-1.621	**-5.429**	**-5.946**
	(0.008)	*(0.213)*	*(0.305)*	*(0.265)*	*(0.000)*	*(0.000)*
Log (roadway length in miles)	**0.403**	0.248	0.160	0.100	**0.539**	**0.504**
	(0.020)	*(0.161)*	*(0.297)*	*(0.552)*	*(0.000)*	*(0.000)*
Population (*1000)	**0.166**	**0.188**	**0.239**	**0.311**	**0.165**	**0.301**
	(0.030)	*(0.001)*	*(0.000)*	*(0.000)*	*(0.005)*	*(0.000)*
Retail employment (*1000)	0.446	-0.442	**0.256**	**0.587**	**0.477**	0.376
	(0.185)	*(0.268)*	*(0.039)*	*(0.003)*	*(0.003)*	*(0.090)*
Non-retail employment (*1000)	**0.066**	**0.100**	**0.126**	0.001	-0.037	0.029
	(0.030)	*(0.044)*	*(0.050)*	*(0.533)*	*(0.392)*	*(0.697)*
Average household income	-0.003	**-0.012**	**-0.012**	**-0.003**	**-0.002**	**0.001**
(*1000)	*(0.327)*	*(0.001)*	*(0.000)*	*(0.027)*	*(0.009)*	*(0.015)*
Overdispersion	**0.263**	**0.178**	**0.264**	**0.338**	**0.381**	**0.175**
	(0.000)	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*	*(0.000)*
Deviance/DF	**0.719**	**0.784**	**0.783**	**0.749**	**0.912**	**0.701**

Notes: first row is the coefficient, second row is the p-significance, and bold coefficients are statistically significant at 5% level of significance.

3.5 APPLICATIONS FOR NETWORK SCREENING

To apply these models, we predicted the number of crashes using the cluster-based SPFs, and estimated the expected number of crashes if no countermeasure had been implemented in the future using the Empirical Bayes (EB) method as prescribed in the HSM (1) for all TAZs in the State. The EB method increases the precision of predictions for the future when only limited historical crash data are available, and it corrects for the regression-to-mean bias (42). Details about our procedures for applying the EB method and developing the network screening application tool are provided in the Appendix to the project final report (31). The resulting EB Expected Crash Counts are added to a GIS layer along with the other data for each TAZ. The resulting GIS layer can be used for reporting and manipulation within a GIS environment by road safety analysts in CTDOT (Connecticut Department of Transportation) and regional or local government to identify locations that have promise for implementing road safety interventions according to HSM procedures (1).

3.6 CONCLUSIONS AND FUTURE RESEARCH

This study demonstrates an alternative for predicting the number of crashes on city or town roads where the traffic volumes are not available. Both intersection SPFs and segment SPFs were estimated at the TAZ level. The TAZs were categorized into six clusters based on land cover intensities and population density using the K-means clustering approach. Cluster-based SPFs were estimated for predicting city and town road intersection and segment crash counts using, respectively, the number of city and town road intersections and the total city and town roadway length. Demographic variables such as population, retail and non-retail employment, total households, and average household income were used as covariates to predict the crash counts.

Due to the high correlation between population and the number of households, two cluster-based SPFs including either population or the number of households were estimated for both intersection and segment crashes. Additionally, an aggregate function using the entire dataset was also developed for comparison. Based on the goodness-of-fit (AIC and BIC values) and prediction performances (MAD and MSPE values), the cluster-based SPFs outperform the aggregate SPFs. The cluster-based SPFs with population perform better than those with the number of households for both intersection and segment crashes.

Finally, the cluster-based SPFs were applied and adjusted using the EB method to produce expected annual crash counts for all TAZs in the State. It is anticipated that the example applications can help regional and municipal agencies identify areas of cities and towns with higher potential for safety improvements, and develop cost-effective countermeasures to improve safety for city and town roads.

This study has demonstrated an initial exploration into developing TAZ level SPFs using demographic variables for city and town roads when the traffic volumes are not available, by clustering TAZs into different types to account for the data heterogeneity. These cluster based TAZ level SPFs can be used to predict the average annual intersection and segment crashes in a TAZ in the context of HSM analyses. They also might be used to help agencies evaluate alternative options for future roadway network and economic development, by identifying the effects of roadway geometric and socio-economic factors on crash counts. However, it is likely to be more difficult to transfer these models to other jurisdictions compared with facility level SPFs (*e.g.* roadway segment and intersection). These TAZ level SPFs are highly dependent upon

not only the clustering of the TAZs, but also the definitions of the TAZs themselves, as well as the character of land development. The relationship between these factors and crash occurrence is likely to vary much more from one place to another than would the relationship between road characteristics and traffic volume. As a consequence, attempts to calibrate these models to another State are not likely to be successful. To use the cluster based TAZ level SPFs, we recommend users to collect their own data and estimate their own SPFs.

One significant challenge in conducting this study was to geo-locate crashes on city and town roads, as at the time of data collection the Connecticut crash data set included only route and milepost. Having geocoded crash records would substantially simplify the process. Other relevant variables that were not available when conducting this study (*e.g.* trip distance and trip duration for a TAZ) may also affect roadway safety, as crash counts are expected to increase with the increase of trip distance and duration in a TAZ. Future research could focus on collecting these variables at the TAZ level, and then estimate new SPFs to improve prediction accuracy.

It is also noted that the observed, predicted and expected annual crash counts for many TAZs were quite small (less than 3). Because each TAZ contains dozens of road segments and intersections, this indicates that the annual crash counts at each individual segment or intersection would be so small as to preclude successful estimation of crash prediction models by segment or intersection. This data condition is further justification for using an area based approach for predicting crashes on city and town roads.

3.7 ACKNOWLEDGMENTS

This research was sponsored by the Joint Highway Research Advisory Council of the University of Connecticut and the Connecticut Department of Transportation through Project 14-1 of the Connecticut Cooperative Transportation Research Program. The contents reflect the views of the authors who are responsible for the accuracy of the information presented herein. The contents do not necessarily reflect the official views or policies of the University of Connecticut or the Connecticut Department of Transportation. The authors would like to thank Mrs. Judy B. Raymond of the Connecticut Department of Transportation for kindly providing demographic data to support this effort. This paper was peer-reviewed by the Transportation Research Board and presented at the 95th annual meeting of the Transportation Research Board, January 2016, Washington, D.C. The authors would also like to thank all reviewers for providing constructive comments to help us improve the paper.

3.8 REFERENCES

1. Highway Safety Manual (2010), 1st Edition, American Association of State Highway and Transportation Officials, Washington D.C..
2. Jonsson, T., Ivan, J. and Zhang, C. (2007). Crash prediction models for intersections on rural multilane highways: differences by collision type. In Transportation Research Record: Journal of the Transportation Research Board, No. 2019, Transportation Research Board of the National Academies, Washington, D.C., pp. 91-98.
3. Ceifetz, A., Bagdade, J., Nabors, D., Sawyer, M., and Eccles, K.(2012). Developing safety plans: A manual for local rural road owner. Project Report, Project 12-017, Federal Highway Administration (FHWA).

4. Souleyrette, R., Caputcu, M., Cook, D., McDonald, T., Sperry, R. and Hans, Z. (2010). Safety Analysis of Low-Volume Rural Roads in Iowa, Final Report, Project 07-309, Institute for Transportation, Iowa State University.
5. Ivan, J. (2004). New approach for including traffic volumes in crash rate analysis and forecasting. In Transportation Research Record: Journal of the Transportation Research Board, No. 1897, Transportation Research Board of the National Academies, Washington, D.C., pp. 134-141.
6. Vogt, A. (1999). Crash Models For Rural Intersections: Four-Lane by Two-Lane Stop-Controlled and Two-Lane by Two-Lane Signalized US Department of Transportation, Federal Highway Administration Report, FHWA-RD-99-128.
7. Vogt, A. and Bared, J. (1998). Accident Models for Two-Lane Rural Segments and Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 1635, Transportation Research Board of the National Academies, Washington, D.C., pp. 18-29.
8. Oh, J., Washington, S. P., and Choi, K. (2004). Development of Accident Prediction Models for Rural Highway Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 1897, Transportation Research Board of the National Academies, Washington, D.C., pp. 18–27.
9. Bindra S., Ivan, J. and Jonsson, T. (2009), Predicting Segment-Intersection Crashes with Land Development Data. In Transportation Research Record: Journal of the Transportation Research Board, No. 2102, Transportation Research Board of the National Academies, Washington, D.C., pp. 9–17.

10. Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B. and Wets, G. (2012). Application of Different Exposure Measures in Development of Planning-level Zonal Crash Prediction Models. In Transportation Research Record: Journal of the Transportation Research Board, No. 2280, Transportation Research Board of the National Academies, Washington, D.C., pp. 145-153.
11. Pulugurtha S., Duddu, V. R, and Kotagiri, Y. (2004). Traffic Analysis Zone Level Crash Estimation Models Based on Land Use Characteristics. Accident Analysis and Prevention, Vol. 36, No. 6, pp. 973–984.
12. Ladron de Guevara, F., Washington, S. P., and Oh, J. (2004). Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. In Transportation Research Record: Journal of the Transportation Research Board, No. 1897, Transportation Research Board of the National Academies, Washington, D.C., pp. 191–199.
13. Lovegrove, G. R., and Sayed, T. (2006). Macro-level Collision Prediction Models for Evaluating Neighborhood Traffic Safety. Canadian Journal of Civil Engineering. Vol. 33. No. 5. pp 609-621.
14. Lovegrove, G. (2012). Road Safety Planning, New Tools for Sustainable Road Safety and Community Development. AV Akademikerverlag. Berlin.
15. Hadayeghi, A., Shalaby, A., and Persaud, B. (2003). Macro-level Accident Prediction Models for Evaluating the Safety of Urban Transportation System. Transportation Research Board. National Research Council. Washington, D.C..
16. Khondakar, B., Sayed, T.and Lovegrove, G. (2010). Transferability of Community-based Collision Prediction Models for Use in Road Safety Planning Applications. Journal of Transportation Engineering. Vol. 136. No. 10. pp. 871-880.

17. Norland, R. B. and Quddus, M. A. (2004). A Spatially Disaggregate Analysis of Road Casualties in England. Accident Analysis and Prevention, Vol. 36, No. 6, pp. 973–984.
18. Jin, Y., Wang, X. and Chen, X. (2011). Incorporating Road Network Structure into Macro Level Traffic Safety Analysis. American Society of Civil Engineers. pp. 2224-2232.
19. Naderan, A. and Shahi, J. (2010). Aggregate Crash Prediction Models: Introducing Crash Generation Concept. Accident Analysis & Prevention. Vol. 42, No. 1. pp. 339-346.
20. Abdel-Aty, M., Siddiqui, C., Huang, H. and Wang, X. (2011). Integrating Trip and Roadway Characteristics to Manage Safety in Traffic Analysis Zones. Transportation Research Record. No. 2213. pp. 20-28.
21. Washington, S. P., Schalkwyk, V., Mitra, S., Meyer, M., Dumbaugh, E., Zoll, M. (2006). Incorporating Safety into Long-Range Transportation Planning. NCHRP Report 546, National Cooperative Highway Research Program, Transportation Research Board, Washington D.C..
22. Lee, J., Abdel-Aty, M. and Jiang, X. (2015). Multivariate Crash Modeling for Motor Vehicle and Non-motorized Modes at the Macroscopic Level. Accident Analysis and Prevention. No. 78. pp. 146-154.
23. Abdel-Aty, M., Lee, J., Siddiqui, C. and Choi, K. (2013). Geographical Unit Based Analysis in the Context of Transportation Safety Planning. Transportation Research Part A. Vol. 49. pp. 62-75.
24. Levine, N., Kim, K. and Nitz, L. (1995). Spatial Analysis of Honolulu Motor Vehicle Crashes. II. Zonal Generators. Accident Analysis and Prevention. Vol. 27, No. 5. pp. 675-685.
25. Norland, R. B. (2003). Traffic Fatalities and Injuries: The Effect of Changes in Infrastructure and Other Trends. Accident Analysis and Prevention. No. 35. Vol. 4. pp. 599-611.

26. Kim, K., Brunner, I. M., and Yamashita, E. Y. (2006). Influence of Land Use, Population, Employment and Economic Activity on Accidents. Transportation Research Record 1953. pp. 56-64.
27. Aguero-Valverde, J. and Jovanis, P. (2006). Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. Accident Analysis & Prevention. Vol. 38. No. 3. pp. 618-625.
28. Huang, H., Abdel-Aty, M. and Darwiche, A. (2010). County-level Crash Risk Analysis in Florida. Transportation Research Record 2149. pp. 27-37.
29. Norland, R. B. and Quddus, M. A. (2004). Analysis of Pedestrian and Bicycle Casualties with Regional Panel Data. Transportation Research Record 1897. pp. 28-33.
30. United States Census Bureau (2010). https://www.census.gov/geo/maps-data/data/tiger-line.html
31. Ivan J., Burnicki, A., Wang K. and Mamun S. (2016). Improvements to Road Safety Improvement Selection Procedures For Connecticut. Connecticut Cooperative Transportation Research Program, Final Report, Project 14-1.
32. CTPP 2010, Census Transportation Planning Package Database. http://ctpp.transportation.org/Pages/5-Year-Data.aspx
33. National Land Cover Database 2011 (2011). http://www.mrlc.gov/nlcd11_data.php
34. Connecticut Crash Data Repository (2016). http://www.ctcrash.uconn.edu/
35. Depaire B., Wets G., and Vanhoof K (2008). Traffic Accident Segmentation by Means of Latent Class Clustering. Accident Analysis and Prevention. Vol. 40. pp. 1257-1266.
36. Hair L., Anderson, R., Tatham, R. and Black, W. (1998). Multivariate Data Analysis. Prentive Hall.

37. Mohamed M., Saunier, N., Miranda-Moreno, L. and Ukkusuri, S. (2013). A Clustering Regresion Approach: A Comprehensive Injury Severity Analysis of Pedestrian-Vehicle Crashes in New York, US and Montreal, Canada. Safety Science. Vol. 54, pp. 27-37.

38. STATA (2011). Clustering Kmeans and Kmedians. Release 12. A Stata Press, StataCorp LP. College Station, Texas http://www.stata.com/manuals13/mvclusterkmeansandkmedians.pdf.

39. Calinski, T. and Harabasz, J.(1974). A Dendriter Method for Cluster Analysis. Communications in Statistics. Vol. 3, pp. 1-27.

40. Washington, S. P., Karlaftis, M., Mannering, F. L. (2011). Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.

41. Oh. J., Lyon, C., Washington S. P., Persaud, B. and Bared, J. (2003). Validation of FHWA Crash Models for Rural Intersections: Lessons Learned. In Transportation Research Record: Journal of the Transportation Research Board. No. 1840, Transportation Research Board of the National Academies, Washington, D.C., pp. 41-49.

42. Hauer, E., Harwood, D. W., Council, F. M. and Griffith, M. S. (2002). The Empirical Bayes Method for Estimating Safety: A Tutorial. In Transportation Research Record. No. 1784. pp.126-131.

4 MULTIVARIATE POISSON LOGNORMAL MODELING OF CRASHES BY TYPE AND SEVERITY ON RURAL TWO LANE HIGHWAYS

4.1 INTRODUCTION AND MOTIVATION

In the United States, improving roadway safety is a high priority of the transportation agencies at the federal, state and local levels, and motor vehicle crashes bring one of the largest economic and societal losses (1). According the National Highway Traffic Safety Administration (NHTSA) (1), there were 32,675 people killed in motor vehicle crashes in 2014, and the total economic losses are up to $836 billion. Given the importance of roadway safety and the substantial economic losses caused by motor vehicle crashes, there has been increasing interest in developing crash prediction models to estimate motor vehicle crash counts, identify crash contributing factors, and implement effective safety strategies and countermeasures to improve traffic safety.

In the current Highway Safety Manual (2), crash counts are estimated in total, even though the crash patterns may vary by crash type and by crash severity. To account for crash frequency variations among crash types, some studies estimated crash counts separately and independently by crash type (3-10). Similarly, in order to accommodate crash frequency variations among crash severities, crash prediction models have been considered by severity level (11-17). The Poisson regression model is widely used in crash prediction research. The limitation of the Poisson model is that the variance of the data is constrained to be equal to the mean. This constraint might be questionable as the variance of crash data is usually greater than the mean, which is known as over-dispersion (18). To address the over-dispersion issue, the Univariate Poisson (UP)

regression and Negative Binomial (NB) regression models are two commonly used approaches to predict the total crash counts or crash counts by crash type or severity. However, all of these models assume crash counts by crash type or severity to be independent. When crash type or severity counts are considered simultaneously, modeling them independently might be questionable, because the crash counts among different crash types or severities may be correlated, due to the presence of shared unobserved factors across crash types or severities for a specific roadway intersection or segment. Neglecting their correlations might lead to biased variance estimation, and reduce model accuracy (19, 20).

In recent years, the Multivariate Poisson (MVP) regression model has been increasingly used to estimate crash counts simultaneously by severity level, as the MVP model is able to account for potential correlation between crashes among different crash severities. Ma and Kockelman (21) applied a MVP model to estimate crash counts by severity level. They found the crash counts are significantly correlated at different levels of injury severity. However, the MVP model cannot account for the overdispersion issue that is usually observed in crash data. Then the Multivariate Poisson Lognormal (MVPLN) model (22) has been applied over the MVP model to accommodate the overdispersion issue in crash severity estimation (19, 23).

The Bayesian framework using Markov Chain Monte Carlo (MCMC) simulation method for MVPLN model is one of the most popular alternatives to simultaneously estimate crash counts by severity level (19, 24, 25). All of these studies indicated that there is a significant correlation across different crash severity counts. Only a few studies that implement the MVPLN model have been conducted in estimating crash counts by crash type. Lee *et al.* (26) used a MVPLN

model to simultaneously estimate crash counts for motor vehicle crashes, bicycle crashes and pedestrian crashes by traffic analysis zone (TAZ) level in central Florida. The study illustrated that the MVPLN model outperforms the univariate model, and there is a significant correlation across the three crash type counts. Serhiyenko *et al.* (20) used a MVPLN model to estimate freeway crashes by crash type in Connecticut. They verified that the crash counts are correlated among different crash types.

Although these MVPLN models can account for both the overdispersion of crash data and correlation among crash types or crash severities, all of these approaches are computationally challenging and time consuming, especially for large data sets with a dependent variable containing many categories (27). In order to improve the computational time, instead of the MCMC simulation method, Serhiyenko *et al.* (20) developed an Integrated Nested Laplace Approximation (INLA) Bayesian approach (28) to jointly estimate freeway crashes by crash type in Connecticut. They verified that the INLA approach can significantly reduce the model running time compared with the MCMC approach.

Overall, to overcome the computational complexity in the current MVPLN models in estimating crash counts by crash severity, and the shortage of research in simultaneously estimating crash counts by crash type and severity, this paper presents MVPLN models using the INLA Bayesian framework to simultaneously estimate crash counts by both crash type and crash severity, using the crash data collected from rural two-lane highways in the USA states of Minnesota and Washington. Furthermore, Negative Binomial (NB) models and Univariate Poisson Lognormal (UPLN) models were respectively estimated to compare with the MVPLN models.

This paper contributes to the exploration of statistical methodologies in predicting crashes by offering a more accurate model with correct variance estimation for the parameters, and identifies the contributing factors on motor vehicle crashes in the context of current HSM analyses to improve traffic safety. The rest of the paper is organized as follows. Section two presents the framework and estimation approaches for MVPLN, UPLN and NB models. The third section describes the data and the fourth section presents the coefficient estimates for the models. The comparisons of model prediction are described in the fifth section, and concluding remarks are presented in section six.

4.2 METHODOLOGIES

4.2.1 FRAMEWORK AND ESTIMATION FOR MVPLN MODEL

Let $y_i = \left(Y_{1i}, Y_{2i} \dots, Y_{Ji}\right)'$ for $i = 1, 2, \dots, n$ be a *J*-dimensional response vector (which represents *J* crash types or crash severities) of crash counts across all *n* intersections/segments on rural two-lane highways. In the MVPLN model, we assume these crash type counts or crash severity counts are correlated, and they are modeled simultaneously with the estimation approach as follows (20).

$$Y_{ji}|\lambda_{ji} \sim Poisson(\lambda_{ji}) \quad (4\text{-}1)$$

where $j = 1, 2, \dots, J$ and $i = 1, 2, \dots, n$ and λ_{ji} represents the mean of the Poisson distribution, which is assumed to be random and can be estimated as:

$$ln(\lambda_{ji}) = \Omega + \boldsymbol{z}'_{ji}\boldsymbol{\beta}_j + \gamma_{ji} \quad (4\text{-}2)$$

where Ω is an offset which represents log exposure for total observation days in the data set for intersection models (*i.e.* in this study, the offset for 3ST, 4ST intersections = log(365*7)=7.85, the offset for 4SG intersections = log(365*6)=7.69), and total observation days times segment

length for segment models (*i.e.* offset for segments = log(365*5*segment length)). $\boldsymbol{z}_{ji}$ represents a vector of covariates, and $\boldsymbol{\beta}_j$ is a vector of coefficients to be estimated. γ_{ji} is a random effect. We assume $\boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i} \dots, \gamma_{Ji})'$ represents a vector of random effects at intersection/segment *i*, and it follows a *J*-dimensional normal distribution, *i.e.*,

$$\boldsymbol{\gamma}_i \sim Normal(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4-3}$$

where **0** is a ***J***-dimensional zero vector, and **Σ** is a ***J*** *****J*** variance-covariance matrix. Let $\boldsymbol{\Sigma}=(\sigma_{rs})_{1\le r\le s\le J}$, and the expectation, variance and covariance for the multivariate crash counts by crash type or severity *r* and *s* can be written as (22, 29):

$$E[Y_{ji}] = exp(\Omega + \boldsymbol{z}'_{ji}\boldsymbol{\beta}_j)exp(\tfrac{\sigma_{jj}}{2}) \tag{4-4}$$

$$Var[Y_{ji}] = exp(offset + \boldsymbol{z}'_{ji}\boldsymbol{\beta}_j)\, exp\left(\tfrac{\sigma_{jj}}{2}\right) + exp\big(2(offset + \boldsymbol{z}'_{ji}\boldsymbol{\beta}_j)\,(exp^2(\sigma_{jj}) - exp(\sigma_{jj})\big) \tag{4-5}$$

$$Cov[Y_{ri}, Y_{si}] = exp(offset + \boldsymbol{z}'_{ji}\boldsymbol{\beta}_j)exp(\tfrac{\sigma_{rr}}{2})exp(\tfrac{\sigma_{ss}}{2})(\exp(\sigma_{rs}) - 1) \tag{4-6}$$

Overdispersion can be accommodated by the MVPLN model, because from equations (4-4) and (4-5), we see that $Var[Y_{ji}] > E[Y_{ji}]$ since the diagonal element of **Σ,** *i.e.* $\sigma_{jj} > 0$. Furthermore, the MVPLN model can also account for the dependence among the components of response vector through equation (4-6), which represents the dependence among crash type counts or crash severity counts. σ_{rs} is the off-diagonal element of the variance-covariance matrix **Σ,** which determines whether the dependence between the r^{th} and s^{th} components of the response vector is positive or negative, and can be written as follows,

$$\sigma_{rs} = \rho_{rs}\sqrt{\sigma_{ss} * \sigma_{rr}} \tag{4-7}$$

where σ_{ss} and σ_{rr} are diagonal elements of the variance-covariance matrix **Σ,** and ρ_{rs} represents a correlation coefficient between -1 and 1 that will be estimated.

Given the functions above, the marginal distribution of the observed crash counts $\boldsymbol{y}_i$ can be derived as:

$$g\left(\boldsymbol{y}_i \middle| \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}\right) = \int \dots \int f_{Normal,J}(\boldsymbol{\gamma}_i|\boldsymbol{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{J} f_{Poisson}\left(y_{ij} \middle| \gamma_{ji}, \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j\right) d\boldsymbol{\gamma}_i \quad (4\text{-}8)$$

where $f_{Normal,J}$ is a *J*-dimensional normal probability density function, and $f_{Poisson}$ is a Poisson probability function. The marginal distribution of the observed crash counts $\boldsymbol{y}_i$ cannot be directly derived, as there is no closed algebraic solution to the *J*-dimensional integral (20). As described earlier, the Markov Chain Monte Carlo (MCMC) simulation approach is usually used to carry out fully Bayesian inference on the model coefficients. However, the MCMC simulation method is computationally challenging and time consuming, especially for big data sets with a high dimensional dependent variable (27). Therefore, a faster Bayesian inference approach is required, and the Integrated Nested Laplace Approximation (INLA) approach proposed by Rue *et al.* (28) and the R-INLA package (30) are applied. A detailed discussion of the INLA Bayesian approach is available in Rue *et al.* (28) and Serhiyenko *et al.* (20).

4.2.2 FRAMEWORK AND ESTIMATION FOR THE UPLN MODEL

The Univariate Poisson Lognormal (UPLN) model can be derived when the dependence between crash counts are ignored (20). Let $Y_{ji}$for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, J$ represent the *J*-dimensional independent crash type or crash severity counts, at each intersection or segment i. Equations (4-1) and (4-2) still hold for the UPLN model, except that in equation (4-2), we assume that γ_{ji} are independent for *j*=1,…, *J*, and

$$\gamma_{ji} \sim Normal(0, {\tau_j}^2) \quad (4\text{-}9)$$

Therefore, the mean and variance of each crash count by crash type or crash severity can be respectively written as:

$$E[Y_{ji}] = exp(\Omega + \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j)exp(\frac{{\tau_j}^2}{2}) \quad (4\text{-}10)$$

$$Var[Y_{ji}] = exp(offset + \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j)\, exp\left(\frac{{\tau_j}^2}{2}\right) + exp\left(2(offset + \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j)\,(exp^2({\tau_j}^2) - exp({\tau_j}^2)\right) \quad (4\text{-}11)$$

The overdispersion can still be accommodated by the UPLN model, because it is illustrated in equation (4-10) and (11), $Var[Y_{ji}] > E[Y_{ji}]$ since ${\tau_j}^2 > 0$. However, there is no covariance term included in the model structure, therefore, the response variable which represents crash counts by crash type or severity in the UPLN model is assumed to be independent. Similar to the MVPLN model, the marginal distribution of the observed crash counts $\boldsymbol{y}_i$ in the UPLN model can be derived as:

$$g(\boldsymbol{y}_i|\boldsymbol{z'}_{ji}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}) = \int \dots \int f_{Normal}(\boldsymbol{\gamma}_i|\boldsymbol{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{J} f_{Poisson}(y_{ij}|\gamma_{ji}, \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j) d\boldsymbol{\gamma}_i \quad (4\text{-}12)$$

The functions and parameters are same as equation (4-8). The INLA Bayesian approach is used to estimate parameters in the UPLN model, and a detailed discussion is available in Rue *et al.* (28) and Serhiyenko *et al.* (20).

4.2.3 FRAMEWORK AND ESTIMATION FOR NB Model

The equation for the Negative Binomial (NB) model can be written as (18):

$$Y_{ji}|\lambda_{ji}, \theta_j \sim NB(\lambda_{ji}, \theta_j) \quad (4\text{-}13)$$

where λ_{ji} is the mean and $\theta_j (\theta_j > 0)$ is the overdispersion factor for the NB model. The mean can be written as:

$$E[Y_{ji}] = ln(\lambda_{ji}) = offset + \boldsymbol{z'}_{ji}\boldsymbol{\beta}_j \quad (4\text{-}14)$$

the offset, $\boldsymbol{z'}_{ji}$ and $\boldsymbol{\beta}_j$ defined in equation (4-2), and the variance can be written as:

$$Var[Y_{ji}] = \lambda_{ji}(1 + \theta_j\lambda_{ji}) = \lambda_{ji} + \theta_j{\lambda_{ji}}^2 \quad (4\text{-}15)$$

the overdispersion can be accommodated by the NB model, because $Var[Y_{ji}] > E[Y_{ji}]$ since $\theta_j > 0$. The maximum likelihood estimation approach and R package (30) are used to estimate the NB parameters.

4.3 DATA PREPARATION AND ANALYSES

In this study, crash data for rural two-lane highways collected from the Highway Safety Information System (HSIS) was used. Considering the data availability, seven-year (2003-2009) crash data for three-way stop controlled (3ST) intersections (n=755 intersections), four-way stop controlled (4ST) intersections (n=1064 intersections), and six-year (2003-2008) crash data for four-way signalized (4SG) intersections (n=63 intersections) were respectively collected from the State of Minnesota. Five-year (2008-2012) crash data for rural two-lane segments (n=7583 segments) was collected from the State of Washington. To obtain sufficient observations in each crash severity and crash type level, crash severity counts were aggregated into three categories (31, 32): 1) PDO (property damage only); 2) B+C which combines type B (possible injury) and type C (non-incapacitating injury) injuries, and 3) K+A which combines type K (fatal) and type A (incapacitating injury) injuries. Crash type counts were aggregated into four categories, based on the original travel direction of involved vehicles (4): 1) Same-direction crashes (SDC) which includes turning-same direction crashes, sideswipe-same direction crashes and rear-end crashes; 2) Intersecting-direction crashes (IDC) which includes turning-intersecting crashes and angle crashes; 3) Opposite-direction crashes (ODC) which includes turning-opposite direction crashes, sideswipe-opposite direction crashes and head-on crashes, and 4) Single-vehicle crashes (SVC) which includes fixed object crashes, jackknife crashes, and run off the road crashes. Note that no IDC crashes are included in rural two-lane segment data, so only SDC, ODC and SVC crashes were estimated in crash type models for segments.

Annual Average Daily Traffic (AADT) for both major and minor roads, and three dummy variables, *i.e.* presence of lighting, presence of left-turn lane and presence of right-turn lane were respectively collected and used as predictors in the intersection models. AADT and three categorical variables, *i.e.* lane width, shoulder width and speed limit were used as predictors in segment models. Furthermore, to verify the model transferability, we also collected six-year (2008-2013) crash data for rural two-lane highways in the State of Connecticut from the Connecticut Crash Data Repository (33). These data include 385 3ST intersections, 61 4ST intersections, 102 4SG intersections and 200 segments and have the same variables as the HSIS data. The explanations and descriptive characteristics for the crash data are summarized in Table 4.1 through Table 4.4.

TABLE 4.1 Descriptive Characteristics of 3ST Intersection Data

	HSIS Data				Connecticut Data			
Continuous Variables	**Min.**	**Max.**	**Mean**	**Std. dev.**	**Min.**	**Max.**	**Mean**	**Std. dev.**
Same-direction Crashes	0.0	19.0	0.6	1.4	0.0	44.0	1.3	3.1
Intersecting-direction Crashes	0.0	11.0	0.3	0.9	0.0	6.0	0.6	1.1
Opposite-direction Crashes	0.0	7.0	0.4	0.8	0.0	5.0	0.3	0.7
Single-vehicle Crashes	0.0	15.0	1.3	1.8	0.0	6.0	0.5	0.8
PDO Crashes	0.0	22.0	1.6	2.5	0.0	40.0	1.8	3.2
B+C Crashes	0.0	15.0	0.8	1.5	0.0	10.0	0.6	1.3
K+A Crashes	0.0	2.0	0.1	0.3	0.0	2.0	0.1	0.2
AADT-Major ($*10^3$)	0.3	21.3	3.6	3.4	0.2	19.6	5.5	4.0
AADT-Minor ($*10^3$)	0.0	6.4	0.6	0.7	0.0	9.9	1.9	1.7
Categorical Variables	**Levels**		**Freq.**	**% of Total**	**Freq.**		**% of Total**	
Presence of Lighting	0		611	80.9%	150		39.0%	
	1		144	19.1%	235		61.0%	
Presence of Left-turn Lane	0		569	75.4%	368		95.6%	
	1		186	24.6%	17		4.4%	
Presence of Right-turn Lane	0		382	50.6%	374		97.1%	
	1		373	49.4%	11		2.9%	

TABLE 4.2 Descriptive Characteristics of 4ST Intersection Data

	HSIS Data				Connecticut Data			
Continuous Variables	**Min.**	**Max.**	**Mean**	**Std. dev.**	**Min.**	**Max.**	**Mean**	**Std. dev.**
Same-direction Crashes	0.0	16.0	0.8	1.5	0.0	11.0	1.5	2.0
Intersecting-direction Crashes	0.0	28.0	1.1	2.1	0.0	8.0	1.5	2.1
Opposite-direction Crashes	0.0	5.0	0.4	0.8	0.0	3.0	0.4	0.7
Single-vehicle Crashes	0.0	11.0	0.9	1.3	0.0	4.0	0.3	0.8
PDO Crashes	0.0	23.0	2.0	2.6	0.0	10.0	2.6	2.8
B+C Crashes	0.0	21.0	1.1	1.8	0.0	9.0	1.0	1.5
K+A Crashes	0.0	4.0	0.1	0.4	0.0	2.0	0.1	0.4
AADT-Major ($*10^3$)	0.1	19.7	3.0	2.2	0.6	15.0	5.3	3.5
AADT-Minor ($*10^3$)	0.0	6.9	0.7	0.8	0.0	5.3	1.5	1.0
Categorical Variables	**Levels**		**Freq.**	**% of Total**	**Freq.**		**% of Total**	
Presence of Lighting	0		822	77.3%	23		37.7%	
	1		242	22.7%	38		62.3%	
Presence of Left-turn Lane	0		1032	97.0%	57		93.4%	
	1		32	3.0%	4		6.6%	
Presence of Right-turn Lane	0		434	40.8%	60		98.4%	
	1		630	59.2%	1		1.6%	

TABLE 4.3 Descriptive Characteristics of 4SG Intersection Data

	HSIS Data				Connecticut Data			
Continuous Variables	**Min.**	**Max.**	**Mean**	**Std. dev.**	**Min.**	**Max.**	**Mean**	**Std. dev.**
Same-direction Crashes	0.0	20.0	3.5	3.7	0.0	32.0	5.5	5.4
Intersecting-direction Crashes	0.0	11.0	2.2	2.5	0.0	18.0	2.0	2.6
Opposite-direction Crashes	0.0	12.0	1.4	2.0	0.0	13.0	1.7	2.2
Single-vehicle Crashes	0.0	8.0	1.0	1.5	0.0	4.0	0.7	1.0
PDO Crashes	0.0	28.0	6.0	5.7	0.0	30.0	7.2	6.4
B+C Crashes	0.0	17.0	2.3	2.8	0.0	12.0	2.5	2.6
K+A Crashes	0.0	1.0	0.1	0.4	0.0	2.0	0.2	0.5
AADT-Major ($*10^3$)	2.1	19.5	9.2	3.9	2.0	26.0	10.5	4.3
AADT-Minor ($*10^3$)	0.3	10.7	4.0	2.6	1.1	12.4	4.8	2.4
Categorical Variables	**Levels**		**Freq.**	**% of Total**	**Freq.**		**% of Total**	
Presence of Lighting	0		4	6.3%	24		23.5%	
	1		59	93.7%	78		76.5%	
Presence of Left-turn Lane	0		41	65.1%	58		56.9%	
	1		22	34.9%	44		43.1%	
Presence of Right-turn Lane	0		29	46.0%	77		75.5%	
	1		34	54.0%	25		24.5%	

TABLE 4.4 Descriptive Characteristics of Segment Data

	HSIS Data				Connecticut Data			
Continuous Variables	**Min.**	**Max.**	**Mean**	**Std. dev.**	**Min.**	**Max.**	**Mean**	**Std. dev.**
Same-direction Crashes	0.0	56.0	0.5	1.9	0.0	15.0	1.8	2.5
Opposite-direction Crashes	0.0	39.0	0.5	1.4	0.0	4.0	0.4	0.7
Single-vehicle Crashes	0.0	14.5	1.9	3.6	0.0	11.0	1.0	1.6
PDO Crashes	0.0	12.7	1.8	3.5	0.0	20.0	3.2	3.3
B+C Crashes	0.0	57.0	1.0	2.1	0.0	6.0	0.6	1.1
K+A Crashes	0.0	11.0	0.2	0.5	0.0	3.0	0.1	0.4
AADT ($*10^3$)	0.0	25.0	3.5	3.6	0.5	19.6	8.5	4.2
Segment Length (in mile)	0.1	7.5	0.5	0.6	0.1	0.9	0.2	0.1
Categorical Variables	**Levels**		**Freq.**	**% of Total**	**Freq.**		**% of Total**	
Lane Width	10ft-14ft		7428	98.0%	198		99.0%	
	<10ft		27	0.3%	2		1.0%	
	>14ft		128	1.7%	0		0.0%	
Shoulder Width	4ft-8ft		4352	57.4%	65		32.5%	
	<4ft		2881	38.0%	126		63.0%	
	>8ft		350	4.6%	9		4.5%	
Speed Limit	<=50mph		1455	19.2%	197		98.5%	
	>50mph		6128	80.8%	3		1.5%	

4.4 DISCUSSION OF RESULTS

Table 4.5 through Table 4.12 show model estimates for the NB, UPLN and MVPLN models by crash type and crash severity for 3ST, 4ST and 4SG intersections and rural two-lane segments. The upper part shows the model parameters, and the lower part shows the correlation coefficients from the MVPLN models. In each table cell, the first row is the coefficient estimate, and the second row is the standard error. Coefficients in boldface are statistically significant at the 5% level, and coefficients with "*" are statistically significant at the 10% level. The coefficient estimates and the standard errors for crash types and crash severities are very similar across the three models. However, standard errors in the MVPLN models are slightly lower than the other two models for some variables, especially when the variables are statistically significant at the 5% level (*e.g.* Major AADT, Minor AADT). Although the difference is extremely small, it still indicates that the MVPLN model can explain extra variations by accounting for the correlation among crash types or crash severities, and lead to more accurate coefficient estimates (20).

Table 4.5 and Table 4.6 respectively show the coefficient estimates for crash types and crash severities at 3ST intersections. Regarding the correlation coefficient from the MVPLN model, the crashes are highly correlated among all crash types and among all crash severities, especially for the correlation between opposite-direction crashes and single-vehicle crashes, in which the correlation coefficient is up to 0.9. As is described in the upper part of the tables, the coefficient estimates for most variables are very close across the three models. The difference can be identified for the variable-presence of lighting for the opposite-direction crashes. In addition, for the intersecting-direction crashes, presence of right turn lane is not significant in MVPLN and UPLN models, but it is significant in the NB model at the 10% level. This verifies the finding in the study conducted by Serhiyenko *et al.* (20) that ignoring the correlation can lead insignificant

parameters to be significant and vice versa. In terms of the effects of variables on crashes, the traffic volume is the most significant and important variable in estimating crashes by crash type and crash severity, and both major AADT and minor AADT have a positively relationship with all crash type and crash severity counts. Presence of lighting is associated with decreased crash numbers for single-vehicle crashes, and for all crash severities.

TABLE 4.5 Estimated Crash Type Models for MN 3ST Intersections (N=755 Intersections)

Coefficient Estimates

Variables	Same-direction Crashes			Intersecting-direction Crashes			Opposite-direction Crashes			Single-vehicle Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-22.25**	**-22.38**	**-21.79**	**-19.91**	**-19.39**	**-18.64**	**-17.31**	**-16.73**	**-16.37**	**-12.78**	**-12.54**	**-12.12**
	0.82	0.84	0.81	0.93	0.93	0.91	0.87	0.83	0.83	0.49	0.48	0.48
Log (Major AADT)	**1.32**	**1.32**	**1.30**	**0.77**	**0.69**	**0.66**	**0.65**	**0.60**	**0.60**	**0.46**	**0.44**	**0.43**
	0.08	0.09	0.09	0.10	0.11	0.11	0.10	0.10	0.10	0.06	0.06	0.06
Log (Minor AADT)	**0.44**	**0.46**	**0.44**	**0.70**	**0.71**	**0.68**	**0.43**	**0.42**	**0.42**	**0.21**	**0.21**	**0.20**
	0.05	0.06	0.06	0.07	0.08	0.08	0.07	0.07	0.07	0.04	0.04	0.04
Presence of Lighting	-0.17	-0.15	-0.15	-0.35	-0.20	-0.10	**-0.41**	-0.22	-0.21	**-0.82**	**-0.80**	**-0.74**
	0.15	0.16	0.16	0.19	0.19	0.19	0.20	0.19	0.19	0.13	0.14	0.14
Presence of Left Turn Lane	-0.16	-0.20	-0.24	-0.30	-0.31	**-0.38**	0.27	0.26	0.24	0.15	0.13	0.14
	0.16	0.17	0.17	0.19	0.20	0.19	0.18	0.18	0.18	0.11	0.12	0.12
Presence of Right Turn Lane	-0.01	0.04	0.01	0.26	0.24	0.33*	0.11	0.16	0.17	0.03	0.03	0.05
	0.14	0.15	0.15	0.18	0.18	0.18	0.18	0.18	0.18	0.11	0.11	0.11
Overdispersion	**0.59**	**0.63**	**0.75**	**0.71**	**0.68**	**0.85**	**0.86**	**0.63**	**0.91**	**0.65**	**0.61**	**0.73**

Correlation Coefficients from MVPLN Model

	Same-direction Crashes	Intersecting-direction Crashes	Opposite-direction Crashes	Single-vehicle Crashes
Same-direction Crashes	**1**			
Intersecting-direction Crashes	**0.78 (<0.001)**	**1**		
Opposite-direction Crashes	**0.54 (<0.001)**	**0.79 (<0.001)**	**1**	
Single-vehicle Crashes	**0.56 (<0.001)**	**0.80 (<0.001)**	**0.90 (<0.001)**	**1**

Notes: the first row is the coefficient estimate; the second row is the standard error; bold coefficient is statistically significant at the 5% significance level; coefficient with * is statistically significant at the 10% significance level.

TABLE 4.6 Estimated Crash Severity Models for MN 3ST Intersections (N=755 Intersections)

Coefficient Estimates									
Variables	PDO Crashes			B+C Crashes			K+A Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-15.48**	**-15.51**	**-14.82**	**-16.15**	**-16.10**	**-15.84**	**-16.28**	**-15.99**	**-15.90**
	0.49	0.49	0.49	0.58	0.58	0.58	1.34	1.30	1.30
Log (Major AADT)	**0.73**	**0.74**	**0.69**	**0.70**	**0.68**	**0.69**	**0.45**	**0.42**	**0.42**
	0.05	0.06	0.06	0.06	0.07	0.07	0.16	0.16	0.16
Log (Minor AADT)	**0.32**	**0.31**	**0.31**	**0.37**	**0.39**	**0.38**	**0.39**	**0.40**	**0.40**
	0.04	0.04	0.04	0.04	0.05	0.05	0.11	0.12	0.12
Presence of Lighting	**-0.36**	**-0.34**	**-0.26**	**-0.63**	**-0.57**	**-0.55**	**-1.10**	**-1.02**	**-1.02**
	0.12	0.12	0.12	0.14	0.14	0.15	0.41	0.41	0.41
Presence of Left Turn Lane	0.08	0.09	0.03	-0.10	-0.13	-0.15	-0.16	-0.21	-0.21
	0.11	0.12	0.12	0.13	0.13	0.13	0.31	0.30	0.30
Presence of Right Turn Lane	-0.04	-0.06	0.02	0.08	0.10	0.09	0.22	0.22	0.22
	0.11	0.11	0.11	0.12	0.12	0.12	0.29	0.28	0.28
Overdispersion	**0.60**	**0.60**	**0.70**	**0.52**	**0.49**	**0.60**	**0.33**	**0.35**	**0.45**

Correlation Coefficients from MVPLN Model

	PDO Crashes	B+C Crashes	K+A Crashes
PDO Crashes	**1**		
B+C Crashes	**0.72 (<0.001)**	**1**	
K+A Crashes	**0.73 (<0.001)**	**0.52 (<0.001)**	**1**

Table 4.7 and Table 4.8 respectively present the parameter estimates for crashes by crash type and crash severity at 4ST intersections. The correlation coefficients are all significant among crashes by cash type and crash severity, which indicates the crashes are dependent among different crash types, as well as different crash severity levels. Similar to the 3ST intersections, major AADT and minor AADT are highly significant, and are positively associated with the number of crashes for all crash types and crash severities. Presence of lighting is statistically significant across all three models for single-vehicle crashes and all crash severity levels, but it is only significant in MVPLN model for opposite-direction crashes, and in both UPLN and NB models for same-direction and intersecting-direction crashes. Presence of right turn lane is only

shown to be statistically significant at the 5% level across three models for the most severe crashes, and is associated with more crash counts for K and A crashes.

TABLE 4.7 Estimated Crash Type Models for MN 4ST Intersections (N=1064 Intersections)

Coefficient Estimates

Variables	Same-direction Crashes			Intersecting-direction Crashes			Opposite-direction Crashes			Single-vehicle Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-19.60**	**-19.53**	**-19.35**	**-16.12**	**-16.20**	**-15.74**	**-17.86**	**-17.51**	**-17.53**	**-13.58**	**-13.50**	**-13.32**
	0.59	0.60	0.61	0.54	0.55	0.54	0.74	0.66	0.72	0.47	0.45	0.46
Log (Major AADT)	**1.03**	**1.02**	**1.02**	**0.38**	**0.39**	**0.36**	**0.79**	**0.77**	**0.78**	**0.53**	**0.52**	**0.52**
	0.07	0.08	0.08	0.07	0.07	0.07	0.09	0.08	0.09	0.06	0.06	0.06
Log (Minor AADT)	**0.50**	**0.50**	**0.49**	**0.79**	**0.79**	**0.79**	**0.42**	**0.41**	**0.41**	**0.23**	**0.22**	**0.22**
	0.04	0.05	0.05	0.04	0.05	0.05	0.06	0.05	0.06	0.03	0.04	0.04
Presence of Lighting	-0.34	**-0.33**	**-0.32**	-0.40	**-0.38**	**-0.34**	**-0.20**	-0.15	-0.16	**-0.73**	**-0.73**	**-0.72**
	0.12	0.12	0.12	0.11	0.11	0.12	0.15	0.13	0.14	0.12	0.12	0.12
Presence of Left Turn Lane	-0.25	-0.27	-0.23	0.23	0.22	0.21	-0.05	-0.10	-0.07	-0.07	-0.08	-0.10
	0.18	0.18	0.18	0.18	0.18	0.18	0.24	0.19	0.22	0.19	0.18	0.19
Presence of Right Turn Lane	-0.10	-0.09	-0.09	0.10	0.11	0.14	-0.11	-0.05	-0.06	0.13	0.13	0.12
	0.11	0.11	0.11	0.10	0.10	0.10	0.14	0.12	0.13	0.09	0.09	0.09
Overdispersion	**0.30**	**0.30**	**0.36**	**0.49**	**0.50**	**0.58**	**0.47**	**0.50**	**0.30**	**0.36**	**0.30**	**0.37**

Correlation Coefficients from MVPLN Model

	Same-direction Crashes	Intersecting-direction Crashes	Opposite-direction Crashes	Single-vehicle Crashes
Same-direction Crashes	**1**			
Intersecting-direction Crashes	**0.60 (<0.001)**	**1**		
Opposite-direction Crashes	**0.61 (<0.001)**	**0.69 (<0.001)**	**1**	
Single-vehicle Crashes	**0.48 (<0.001)**	**0.50 (<0.001)**	**0.82 (<0.001)**	**1**

TABLE 4.8 Estimated Crash Severity Models for MN 4ST Intersections (N=1064 Intersections)

Coefficient Estimates									
Variables	PDO Crashes			B+C Crashes			K+A Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-15.71**	**-15.74**	**-15.52**	**-15.56**	**-15.58**	**-15.30**	**-16.45**	**-16.40**	**-16.29**
	0.39	0.39	0.39	0.48	0.48	0.48	1.07	1.03	1.11
Log (Major AADT)	**0.70**	**0.70**	**0.69**	**0.52**	**0.51**	**0.49**	**0.24**	**0.25**	0.23
	0.05	0.05	0.05	0.06	0.06	0.06	0.14	0.13	0.15
Log (Minor AADT)	**0.46**	**0.46**	**0.46**	**0.56**	**0.57**	**0.58**	**0.66**	**0.67**	**0.68**
	0.03	0.03	0.03	0.04	0.04	0.04	0.09	0.09	0.10
Presence of Lighting	**-0.20**	**-0.20**	**-0.20**	**-0.68**	**-0.67**	**-0.67**	**-0.88**	**-0.86**	**-0.85**
	0.08	0.08	0.08	0.09	0.11	0.11	0.28	0.27	0.28
Presence of Left Turn Lane	-0.02	-0.02	-0.04	0.22	0.20	0.23	0.20	0.17	0.11
	0.14	0.14	0.14	0.16	0.16	0.16	0.33	0.30	0.35
Presence of Right Turn Lane	-0.06	-0.05	-0.06	0.13	0.14	0.13	**0.72**	**0.72**	**0.73**
	0.07	0.07	0.07	0.09	0.09	0.09	0.24	0.23	0.25
Overdispersion	**0.29**	**0.28**	**0.31**	**0.36**	**0.34**	**0.38**	**0.41**	**0.43**	**0.63**

Correlation Coefficients from MVPLN Model			
	PDO Crashes	B+C Crashes	K+A Crashes
PDO Crashes	**1**		
B+C Crashes	**0.72 (<0.001)**	**1**	
K+A Crashes	**0.64 (<0.001)**	**0.66 (<0.001)**	**1**

Table 4.9 and table 4.10 provide the parameter estimates for the three models at 4SG intersections. In terms of the correlation coefficients, the crash counts between same-direction crashes and opposite-direction crashes are evidently correlated at the 5% significance level. Single-vehicle crashes and same-direction crashes, opposite-direction crashes and intersecting-direction crashes are statistically significant at 10% level. For the crash severity model, only the PDO crashes and B and C crashes are shown to be positively correlated at the 10% significance level. Due to the low sample size which only contains 63 signalized intersections, most variables are not identified to be significant. It seems traffic volume is still associated with increased crash counts for all crash types and crash severities. Presence of lighting is significant in all three models for intersecting-direction crashes, but it is only significant in UPLN model at the 5%

level, and in NB model at the 10% level for B and C crashes. Presence of left turn lane only has a positive relationship with the opposite-direction crashes in UPLN model. Presence of right turn lane seems to have a positive relationship with crash counts for all crash types and crash severities in all three models, but it is mainly significant for intersecting-direction crashes and B and C crashes.

TABLE 4.9 Estimated Crash Type Models for MN 4SG Intersections (N=63 Intersections)

Coefficient Estimates												
Variables	Same-direction Crashes			Intersecting-direction Crashes			Opposite-direction Crashes			Single-vehicle Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-20.63**	**-20.60**	**-20.41**	**-16.30**	**-16.58**	**-15.93**	**-13.23**	**-13.44**	**-13.44**	**-26.02**	**-26.15**	**-25.97**
	2.64	2.54	2.55	2.84	2.87	2.89	3.26	2.89	3.51	4.21	3.75	4.53
Log (Major AADT)	**1.11**	**1.09**	**1.08**	0.48	0.48	0.50	-0.12	-0.23	-0.08	**1.76**	**1.79**	**1.79**
	0.31	0.30	0.30	0.33	0.33	0.33	0.40	0.34	0.43	0.47	0.41	0.51
Log (Minor AADT)	**0.41**	**0.42**	**0.43**	0.23	**0.25**	0.22	**0.72**	**0.85**	**0.73**	0.06	0.04	0.09
	0.13	0.13	0.13	0.15	0.14	0.14	0.22	0.20	0.23	0.19	0.16	0.20
Presence of Lighting	0.41	0.49	0.49	**2.25**	**2.31**	**2.28**	0.26	0.44	0.30	0.80	0.86	0.86
	0.55	0.53	0.52	1.05	1.05	1.05	0.58	0.49	0.62	1.07	1.04	1.12
Presence of Left Turn Lane	-0.21	-0.15	-0.13	-0.04	0.03	0.03	0.47	**0.60**	0.50	-0.50	-0.41	-0.48
	0.26	0.25	0.24	0.29	0.28	0.28	0.34	0.26	0.35	0.37	0.31	0.39
Presence of Right Turn Lane	0.21	0.24	0.22	**0.73**	**0.72**	**0.72**	0.46	**0.56**	0.47	0.70*	**0.73**	0.63
	0.25	0.24	0.23	0.28	0.28	0.27	0.33	0.27	0.34	0.39	0.34	0.40
Overdispersion	**0.26**	**0.18**	**0.24**	**0.20**	**0.16**	**0.29**	**0.20**	**0.23**	**0.42**	**0.18**	**0.20**	**0.39**

Correlation Coefficients from MVPLN Model

	Same-direction Crashes	Intersecting-direction Crashes	Opposite-direction Crashes	Single-vehicle Crashes
Same-direction Crashes	**1**			
Intersecting-direction Crashes	**0.61 (<0.001)**	**1**		
Opposite-direction Crashes	**0.63 (<0.001)**	0.45* (0.088)	**1**	
Single-vehicle Crashes	0.51* (0.073)	0.27 (0.203)	0.42 (0.104)	**1**

TABLE 4.10 Estimated Crash Severity Models for MN 4SG Intersections (N=63 Intersections)

Coefficient Estimates									
Variables	PDO Crashes			B+C Crashes			K+A Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-18.90**	**-19.46**	**-19.00**	**-17.82**	**-18.71**	**-17.65**	**-40.48**	**-59.79**	-43.47
	2.37	2.39	2.27	2.81	2.37	2.68	11.25	17.13	37.20
Log (Major AADT)	**1.00**	**1.05**	**1.02**	**0.71**	**0.73**	**0.68**	1.43	1.88	1.80
	0.28	0.28	0.27	0.33	0.26	0.31	1.10	1.13	1.13
Log (Minor AADT)	**0.40**	**0.40**	**0.40**	**0.34**	**0.41**	**0.36**	0.01	0.06	0.06
	0.12	0.12	0.12	0.15	0.12	0.14	0.41	0.41	0.41
Presence of Lighting	0.35	0.36	0.42	1.02	**1.13**	1.09*	2.64	4.37	16.29
	0.47	0.46	0.45	0.67	0.60	0.64	6.29	13.76	37.19
Presence of Left Turn Lane	-0.22	-0.20	-0.16	0.18	0.32	0.28	0.53	0.46	0.43
	0.25	0.24	0.23	0.28	0.20	0.24	0.88	0.85	0.85
Presence of Right Turn Lane	0.33	0.32	0.31	**0.62**	**0.63**	**0.64**	0.49	0.46	0.43
	0.23	0.23	0.21	0.28	0.22	0.25	0.99	0.94	0.94
Overdispersion	**0.29**	**0.27**	**0.30**	**0.21**	**0.18**	**0.15**	**0.15**	**0.20**	**0.18**

Correlation Coefficients from MVPLN Model			
	PDO Crashes	B+C Crashes	K+A Crashes
PDO Crashes	**1**		
B+C Crashes	0.60* (0.031)	**1**	
K+A Crashes	0.29 (0.197)	0.28 (0.190)	**1**

Table 4.11 and table 4.12 describe the model parameters for rural two-lane segments. The estimated correlation coefficients in the MVPLN model prove that crash counts are highly correlated among crash types and crash severities, which indicates that it is preferable to accommodate these correlations when crashes are simultaneously estimated by crash type and crash severity. For the parameter estimates, traffic volume is significant in all models, and it is positively associated with crashes for all crash types and crash severities. Lane width, shoulder width and speed limit are three categorical variables. Compared with the reference level for lane width (*i.e.* lane width between 10 ft and 14 ft), narrower lane width is associated with increased single-vehicle crashes, and less severe crashes, including PDO, B and C crashes. Wider lane width is associated with increased opposite-direction crashes, but it is associated with decreased

single-vehicle crashes. This might be because when the driving lane is wider, drivers are more likely to drive faster and cross the center line of highway, which results in more head-on crashes rather than run-off-road crashes. This finding is counterintuitive to the HSM when considering modeling crashes in total, in which wider lane width is uniformly associated with less crash counts. To verify this phenomenon, we also developed a NB model to predict total crashes on rural two-lane segments. The coefficient estimate for wider lane width is -0.07, with the standard error 0.11, which is consistent with the results in the HSM. These different impacts of roadway geometric factors on different crash types or severities cannot be identified when modeling crashes in total, which suggests that estimating crashes by crash type or severity might be more helpful in roadway safety analyses, and in particular for identifying countermeasures. Compared with shoulder width that falls between 4 ft and 8 ft, lower shoulder width is positively associated with crashes for all crash types and crash severities. A wider shoulder is negatively associated with all crash type and crash severity counts, but it is only statistically significant for single-vehicle crashes and B and C crashes. Roadway segments with higher speed limit are associated with decreased crashes for all crash severity levels, and both same-direction and opposite-direction crashes, but are associated with increased single-vehicle crashes. Similar with the impacts of wider lane width, these different impacts cannot be identified when crashes are estimated in total.

TABLE 4.11 Estimated Crash Type Models for WA Rural Two-Lane Segments (N=7583 Segments)

Coefficient Estimates									
Variables	Same-direction Crashes			Opposite-direction Crashes			Single-vehicle Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-21.74**	**-22.00**	**-20.48**	**-17.09**	**-17.40**	**-16.21**	**-12.13**	**-12.21**	**-11.89**
	0.31	0.31	0.31	0.27	0.28	0.27	0.13	0.13	0.13
Log (AADT)	**1.69**	**1.76**	**1.65**	**1.13**	**1.20**	**1.14**	**0.70**	**0.72**	**0.71**
	0.04	0.04	0.03	0.03	0.03	0.03	0.02	0.02	0.02
Lane Width (<10ft)	-5.80	-5.87	-23.05	-0.01	0.04	-0.03	**0.75**	**0.75**	**0.77**
	12.16	12.12	37.66	0.57	0.57	0.55	0.21	0.21	0.22
Lane Width (>14ft)	0.28	0.33	0.28	**0.55**	**0.53**	**0.58**	**-0.47**	**-0.46**	**-0.39**
	0.18	0.19	0.19	0.18	0.19	0.18	0.14	0.15	0.15
Shoulder Width (<4ft)	**0.19**	**0.20**	**0.19**	**0.17**	**0.17**	**0.13**	**0.20**	**0.20**	**0.17**
	0.06	0.06	0.06	0.06	0.06	0.06	0.03	0.03	0.03
Shoulder Width (>8ft)	-0.07	-0.07	-0.14	-0.18	-0.13	-0.13	**-0.14**	**-0.14**	**-0.18**
	0.11	0.12	0.12	0.11	0.12	0.12	0.06	0.06	0.07
Speed Limit (>50mph)	**-0.51**	**-0.62**	**-0.75**	**-0.31**	**-0.48**	**-0.57**	**0.17**	**0.15**	**0.12**
	0.06	0.06	0.06	0.06	0.06	0.06	0.03	0.04	0.04
Overdispersion	**0.90**	**0.99**	**1.25**	**1.01**	**1.16**	**1.59**	**0.46**	**0.47**	**0.57**

Correlation Coefficients from MVPLN Model			
	Same-direction Crashes	Opposite-direction Crashes	Single-vehicle Crashes
Same-direction Crashes	**1**		
Opposite-direction Crashes	**0.75 (<0.001)**	**1**	
Single-vehicle Crashes	**0.38 (<0.001)**	**0.51 (<0.001)**	**1**

TABLE 4.12 Estimated Crash Severity Models for WA Rural Two-Lane Segments (N=7583 Segments)

Coefficient Estimates									
Variables	PDO Crashes			B+C Crashes			K+A Crashes		
	MVPLN	UPLN	NB	MVPLN	UPLN	NB	MVPLN	UPLN	NB
Intercept	**-13.89**	**-14.12**	**-13.70**	**-14.67**	**-14.68**	**-14.09**	**-14.20**	**-14.32**	**-13.91**
	0.14	0.14	0.14	0.17	0.18	0.18	0.32	0.33	0.32
Log (AADT)	**0.93**	**0.97**	**0.96**	**0.94**	**0.95**	**0.93**	**0.66**	**0.70**	**0.69**
	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.04
Lane Width (<10ft)	**0.67**	**0.68**	**0.66**	**0.68**	**0.62**	**0.56**	-0.80	-0.95	-0.97
	0.23	0.24	0.24	0.29	0.30	0.30	1.02	1.02	1.02
Lane Width (>14ft)	-0.03	0.08	0.08	-0.16	-0.02	0.02	-0.01	0.15	0.15
	0.12	0.12	0.12	0.16	0.16	0.16	0.32	0.32	0.32
Shoulder Width (<4ft)	**0.20**	**0.20**	**0.18**	**0.23**	**0.23**	**0.20**	0.11	0.11	0.10
	0.03	0.03	0.03	0.04	0.04	0.04	0.07	0.07	0.07
Shoulder Width (>8ft)	-0.06	-0.04	-0.08	**-0.26**	**-0.31**	**-0.35**	-0.26	-0.30	-0.31*
	0.06	0.06	0.07	0.08	0.09	0.09	0.17	0.18	0.18
Speed Limit (>50mph)	**-0.07**	**-0.13**	**-0.19**	**-0.14**	**-0.11**	**-0.18**	0.04	-0.13	-0.15
	0.03	0.04	0.04	0.04	0.05	0.05	0.09	0.09	0.09
Overdispersion	**0.44**	**0.45**	**0.55**	**0.54**	**0.59**	**0.71**	**0.56**	**0.67**	**0.83**

Correlation Coefficients from MVPLN Model			
	PDO Crashes	B+C Crashes	K+A Crashes
PDO Crashes	**1**		
B+C Crashes	**0.90 (<0.001)**	**1**	
K+A Crashes	**0.79 (<0.001)**	**0.82 (<0.001)**	**1**

4.5 MODEL COMPARISON

In order to evaluate the model prediction ability, we randomly select 80% of the data to estimate the model parameters, and use the remaining 20% hold-out data to compare the model prediction performance based on the Predicted Mean Absolute Error (PMAE). It is noted here that the data separation process is only used in model prediction comparison. The model coefficient estimates in the context use the entire data sets. To account for the possible bias in random selection of samples, we duplicated this process 30 times, and calculated the Average Predicted Mean Absolute Error (APMAE) (20), which can be measured as:

$$APMAE_j = \frac{1}{30}\sum_{l=1}^{L}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_{ji,l} - Y_{ji,l}\right|\right) \quad (4\text{-}16)$$

where j=1, 2, ……, J for crash types or crash severities, i=1, 2, ……, n for intersections or segments, l=1, 2, ……, L for sample selection times, which is equal to 30 in our study. $\hat{Y}_{ji,l}$ is the predicted crash counts for crash type or severity j, at intersection or segment i in the j^{th} sample, and $Y_{ji,l}$ is the observed crash counts for crash type or severity j, at intersection or segment i in the j^{th} sample. Furthermore, in order to evaluate the model transferability, we also collected similar intersection and segment data from the State of Connecticut, and used them to calculate the APMAE values for model comparison as well.

Figure 4.1 shows the prediction performance of MVPLN, UPLN and NB models by crash type using the Minnesota intersection data and Washington segment data. The MVPLN model shows a significant improvement in prediction of crash types at 3ST intersections, especially for the same-direction crashes and single-vehicle crashes. Both MVPLN and UPLN models obviously outperform NB model in terms of all four crash types at 4ST interserctions and rural two-lane segments, in which the MVPLN model performs slightly better than the UPLN model. Compared with the 3ST, 4ST intersections and rural two-lane segments, the APMAE values across the three models for all crash types at 4SG intersections are very close, which might be due to the small sample size of 4SG intersections (63 cases) in Minnesota data. This indicates that large data sample would benefit the MVPLN model in crash prediction performance.

Figure 4.2 presents the model prediction comparison by crash type using the Connecticut data (which were not used to estimate the models). For 3ST intersections and rural two-lane segments, MVPLN model performs slightly better than the UPLN model, but dramatically better than the

NB model. The prediction performance at 4ST and 4SG intersections is similar across the three models.

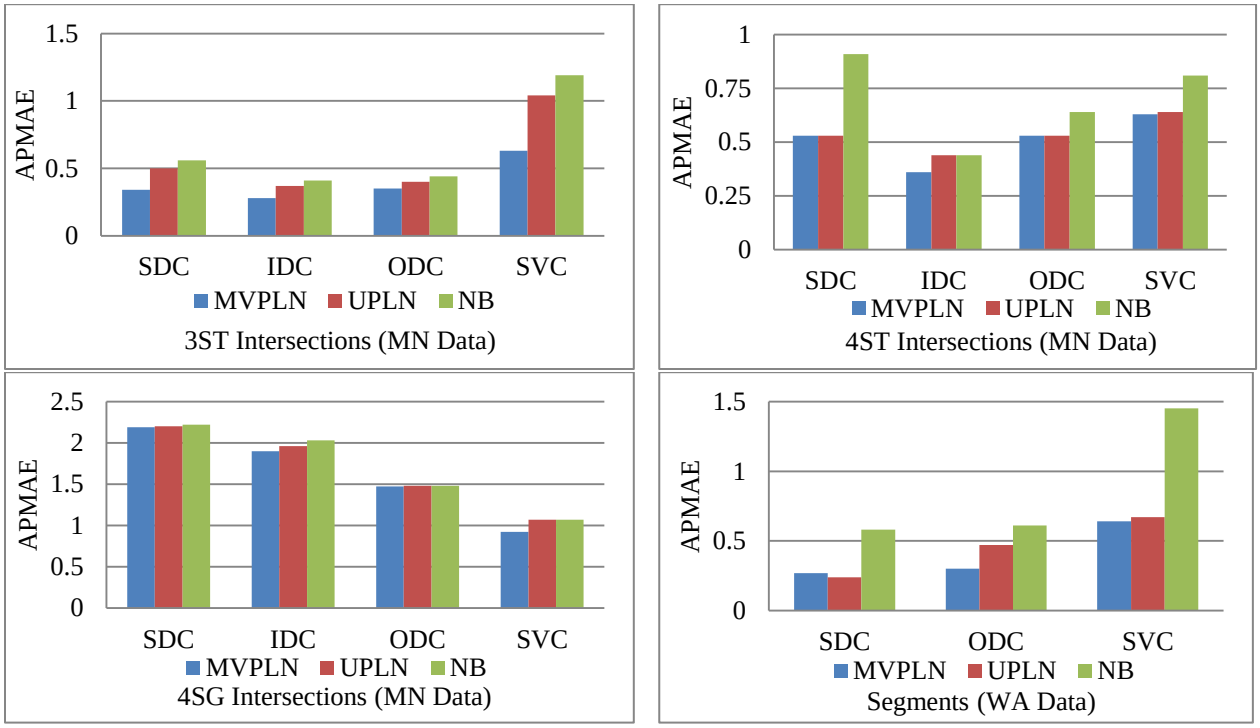


FIGURE 4.1 Prediction Performance of Crash Type Models (MN and WA Data)

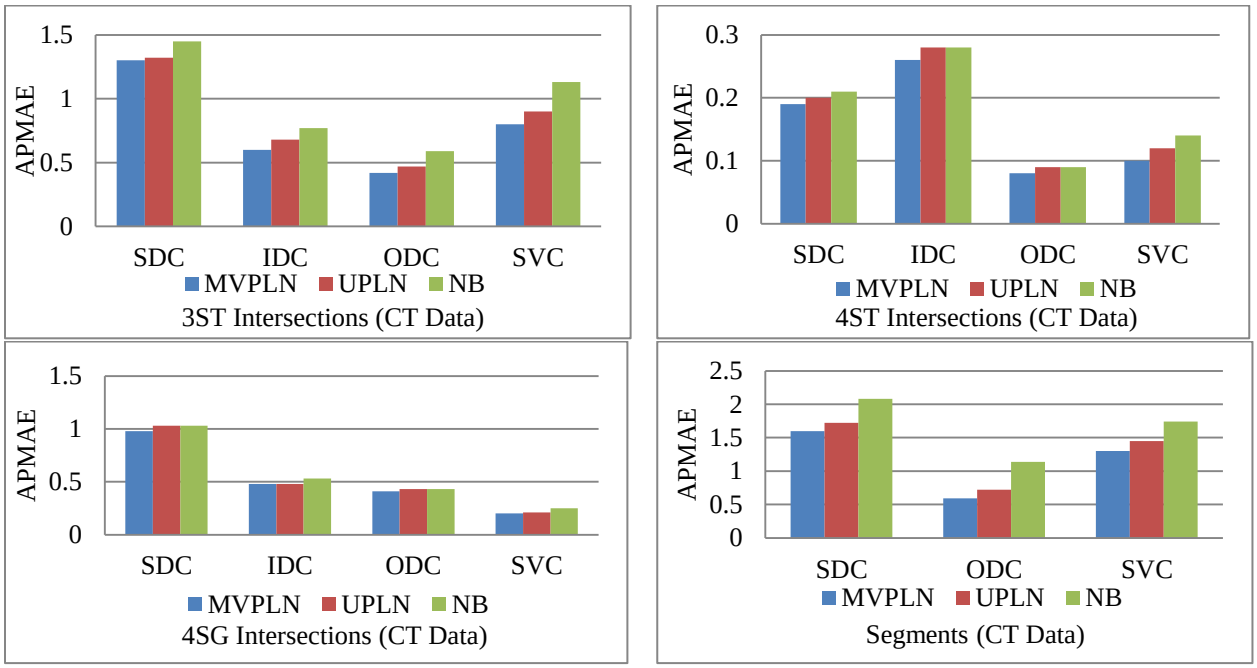


FIGURE 4.2 Prediction Performance of Crash Type Models (CT Data)

Figure 4.3 compares the prediction performance for the three models by crash severity using the Minnesota and Washington data. The MVPLN model is shown to significantly improve the prediction accuracy for all facilities, especially for the prediction of less severe crashes (*i.e.* PDO, C and B crashes). Figure 4.4 shows the model prediction performance using the Connecticut data. The MVPLN model and UPLN model have a similar prediction performance, but both of the two models outperform the NB models in predicting crash severity counts, especially at 3ST intersections and rural two-lane segments.

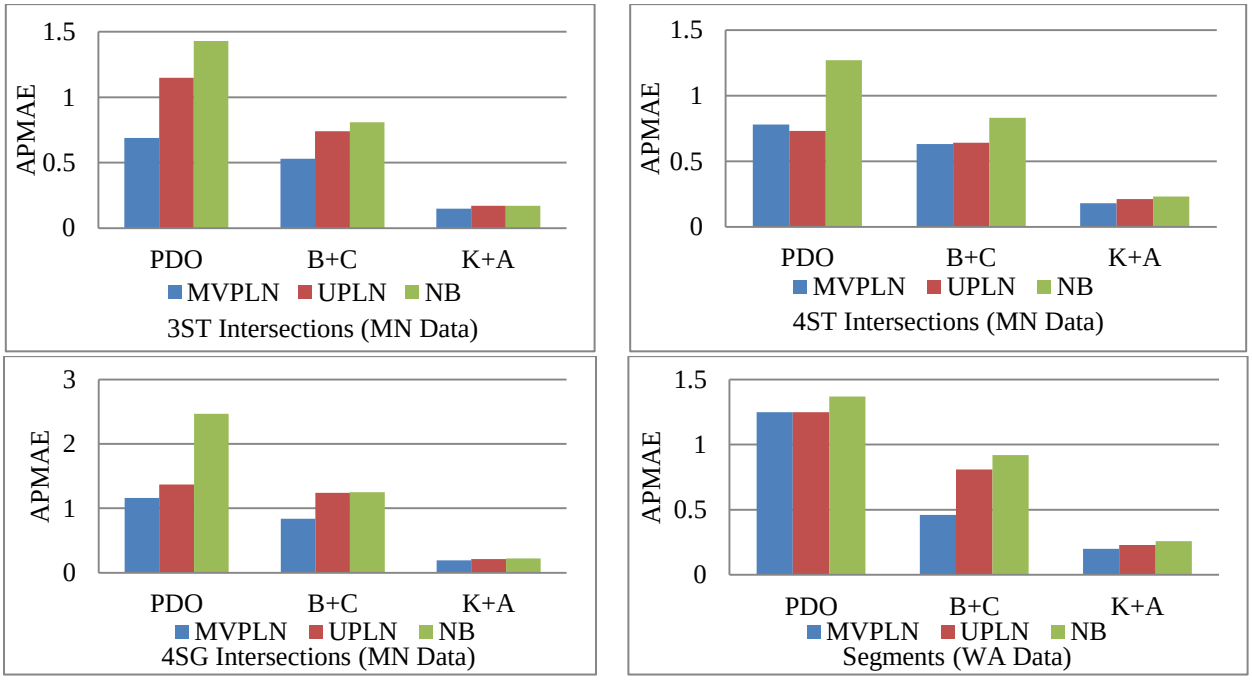


FIGURE 4.3 Prediction Performance of Crash Severity Models (MN and WA Data)

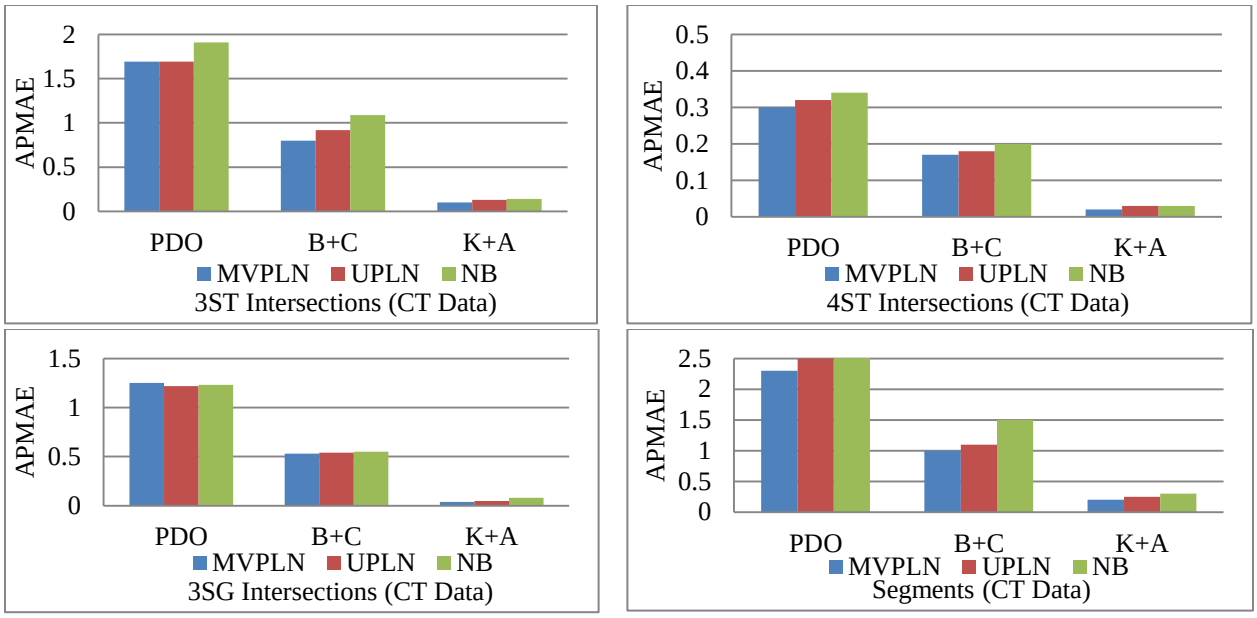


FIGURE 4.4 Prediction Performance of Crash Severity Models (CT Data)

Overall, based on the findings through Figure 4.1 to Figure 4.4, in simultaneously predicting crash counts by crash type and crash severity, the MVPLN model slightly outperforms the UPLN model, while it is significantly better than the NB model. The three models have a close prediction performance in predicting crashes at 4SG intersections, which had the smallest sample size (63 cases), suggesting that the MVPLN model requires a large data sample to achieve its greater accuracy. Considering the data source, the model prediction using the Connecticut data performs as well as using the Minnesota and Washington data, based on the APMAE values, which indicates that the estimated MVPLN models in this study can be transferrable to predict crashes in other States. Moreover, there is strong evidence shown from Table 4.5 to Table 4.12 that the estimated correlation coefficients are statistically significant for all intersection and segment models, which indicates the crash type counts and crash severity counts are highly dependent, and ignoring the correlation would lead to incorrect variance estimation for model parameters, thus we can conclude that the MVPLN model should be considered in predicting

crash counts by crash type and severity over the UPLN and NB models, particularly when all of the models are expected to be used simultaneously.

4.6 DISCUSSION AND CONCLUSION

This paper presents the use of INLA MVPLN model to estimate crashes by crash type and severity at rural two-lane 3ST, 4ST and 4SG intersections in Minnesota State, as well as rural two-lane segments in Washington State. The crash type is categorized into same-direction crashes, intersecting-direction crashes, opposite-direction crashes and single-vehicle crashes. The crash severity is categorized into PDO crashes, B and C crashes, K and A crashes. AADT for both major and minor roads, presence of lighting, presence of left-turn lane and presence of right turn lane were used as predictors in intersection models, and AADT, segment length, lane width, shoulder width and speed limit are used as predictors in segment models.

The coefficient estimates of MVPLN models demonstrate that the traffic volume has an increased effect on crash counts by crash type and severity. Presence of lighting is shown to be associated with decreased crash counts at 3ST, 4ST intersections, but it is associated with increased crash counts at 4SG intersections, although it is not statistically significant. Presence of left-turn lane and right-turn lane are barely significant at all intersection models. In terms of the rural two-lane segments, the effects of most roadway geometric factors on different crash type counts and crash severity counts are consistent, except for segments with wider lane width and higher speed limit. The results show that wider lane width is associated with increased opposite-direction crashes, but it is associated with decreased single-vehicle crashes. Roadway segments with higher speed limit are associated with decreased crashes for same-direction and opposite-direction crashes, but are associated with increased single-vehicle crashes. These different

impacts of roadway geometric factors for different crash types or severities cannot be identified when modeling crashes in total, which suggests that estimating crashes by crash type or severity might be more helpful in identifying crash contributing factors. UPLN models and NB models are respectively developed to compare with the MVPLN models. The parameter estimates are shown to be close across the three models, while the standard errors estimated in MVPLN models are slightly lower than the other two models in most cases. The correlation coefficients in MVPLN models demonstrate that the crash counts are highly correlated among crash types and crash severities, which indicates that the MVPLN model can lead to more accurate variance estimates by accounting for the possible correlations among crash type and severity counts.

In the end, the prediction performance of MVPLN models is compared to the UPLN and NB models based on the APMAE values, by using both Minnesota and Washington data and Connecticut data. The MVPLN model is shown to outperform the UPLN and NB models for crashes at 3ST, 4ST intersections and rural two-lane segments. The models have a close prediction performance in predicting crashes at 4SG intersections, which suggests that the large data sample will benefit the prediction performance of the MVPLN model. The model prediction using the Connecticut data performs as well as using the Minnesota and Washington data, which verifies that the estimated MVPLN models are transferable to be used by other States or agencies. In summary, we conclude that the MVPLN model can lead to correct variance estimation, by accounting for the correlations among crash type and severity counts, and should be considered when simultaneously predicting crash counts by crash type and crash severity for rural two-lane intersections and segments.

It is expected that this research can develop an approach to rigorously estimate crashes by crash type and crash severity in the context of current HSM analyses, and can identify the influence of roadway geometric characteristics on crash counts by crash type and crash severity, to help agencies assess roadway design alternatives, implement roadway safety facilities to reduce traffic crashes, especially the severe ones. One significant challenge in conducting this study is that the data sample for 4SG intersections is too small to compare the MVPLN model with the UPLN and NB models regarding the prediction performance. Future work can focus on collecting more rural two-lane 4SG intersections, and then estimate new MVPLN models to improve prediction accuracy. Future work can also target on estimating MVPLN models by crash type and severity for other roadway facility types, such as rural multilane highways, urban and suburban arterials, and freeways.

4.7 ACKNOWLEDGMENTS

This research was sponsored by the Connecticut Transportation Institute (CTI) of the University of Connecticut. The contents reflect the views of the authors who are responsible for the accuracy of the information presented herein. The contents do not necessarily reflect the official views or policies of the University of Connecticut. The authors would like to thank Dr. Volodymyr Serhiyenko for kindly providing guidance on use of the INLA approach, and the Connecticut Transportation Crash Data Repository (CTCDR) and the Highway Safety Information System (HSIS) for providing the data.

4.8 REFERENCES

1. NHTSA. National Highway Traffic Safety Administration. http://www.nhtsa.gov/NCSA
2. Highway Safety Manual, 1st Edition, 2010. American Association of State Highway and Transportation Officials, Washington D.C.

3. Ivan, J. N., Wang C. and Bernard N. R., 2000. Exploring Two-Lane Highway Crash Rates Using Land Use and Hourly Exposure. Accident Analysis & Prevention. Vol. 32. No. 6. pp. 787-795.
4. Qin, X., 2002. Selecting Exposure Measures for Predicting Crash Rates on Two-lane Rural Highways. Doctoral Dissertations. Paper AAI3066254. http://digitalcommons.uconn.edu/dissertations/AAI3066254
5. Poch, M., and Mannering F. L., 1996. Negative Binomial Analysis of Intersection-Accident Frequencies. Journal of Transportation Engineering. Vol. 122. No. 2. pp. 105-113.
6. Hauer, E., 2000. Observational Before-After Studies in Road Safety: Estimating the Effects of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd.
7. Abdel-Aty, M. A., Keller, J. and Brady, P. A., 2005. Analysis of Types of Crashes at Signalized Intersections by Using Complete Crash Data and Tree-based Regression. Transportation Research Record 1908. pp. 37-45.
8. Hauer, E., Ng, J. and Lovell, J., 1988. Estimation of Safety at Signalized Intersections. Transportation Research Record 1185. pp. 48-61.
9. Shankar, V., Manning, F. and Barfield, W., 1995. Effect of Roadway Geometric and Environmental Factors on Rural Freeway Accident Frequencies. Accident Analysis and Prevention. Vol. 27, No. 3. pp. 371-389.
10. Geedipally, S. R. and Lord, D., 2010. Investing the Effect of Modeling Single-vehicle and Multi-vehicle Crashes Separately on Confidence Intervals of Poisson-Gamma Models. Accident Analysis and Prevention. Vol. 42, No. 4. pp. 1273-1282.
11. Abdel-Aty, M. A and Radwan A. E., 2000. Modeling Traffic Accident Occurrence and Involvement. Accident Analysis and Prevention. Vol. 32. No. 5. pp. 633-642.

12. Ulfarsson, G. F. and Shankar V. N., 2002. Accident Count Model Based on Multiyear Cross-Sectional Roadway Data with Serial Correlation. Transportation Research Record 1840. pp. 193-197.
13. Lord, D. and Persaud B. N., 2000. Accident Prediction Models with and without Trend: Application of the Generalized Estimating Equations Procedure. Transportation Research Record 1717. pp. 102-108.
14. Lord, D., Geedipally, S. R., Persaud, B., Washington, S. P., Schalkwyk, I., Ivan, J., Lyon, C. and Jonsson, T., 2008. NCHRP 126. Methodology to Predict the Safety Performance Function of Rural Multilane Highways. Transportation Research Board of the National Academics, Washington D.C.
15. Lyon C., Oh, J., Persaud, B., Washington, S. and Bared, J., 2003. Empirical Investing of Interactive Highway Safety Design Model Accident Prediction Algorithm: Rural Intersections. Transportation Research Record 1840. pp. 78-86.
16. Tarko, A. P., Inerowicz, M., Ramos, J. and Li, W., 2008. Tool with Road-level Crash Prediction for Transportation Safety Planning. Transportation Research Record 2083. pp. 16-25.
17. Geedipally, S., Patil, S. and Lord, D., 2010. Examination of Methods to Estimate Crash Counts by Collision Type. Transportation Research Record 2165. pp. 12-20.
18. Washington, S., Karlaftis, M. and Mannering F. L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.
19. Ma, J., Kockelman K. and Damien P., 2008. A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods. Accident Analysis & Prevention. Vol. 40, No. 3. pp. 964-975.

20. Serhiyenko, V., Mamun, S., Ivan J. and Ravishanker N., 2016. Fast Bayesian Inference for Modeling Multivariate Crash Counts. Analytic Methods in Accident Research.
21. Ma, J. and Kockelman K., 2006. Bayesian Multivariate Poisson-Lognormal Regression for Models of Injury Count by Severity. Transportation Research Record 1950. pp. 24-34.
22. Chib, S. and Winkelmann, R., 2001. Markov Chain Monte Carlo Analysis of Correlated Count Data. Journal of Business and Economic Statistics, Vol. 19, No 4. pp. 428-435.
23. Park, E. S. and Lord, D., 2007. Multivariate Poisson-lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record 2019. pp. 1-6.
24. Aguero-Valverde, J. and Jovanis P. P., 2009. Bayesian Multivariate Poisson Log-Normal Models for Crash Severity Modeling and Site Ranking. Presented at the 88th Annual Meeting of the Transportation Research Board.
25. El-Basyouny, K. and Sayed T., 2009. Collision Prediction Models using Multivariate Poisson-Lognormal Regression. Accident Analysis and Prevention. Vol. 41. No. 4. pp. 820-828.
26. Lee, J., Abdel-Aty, M. and Jiang, X., 2015. Multivariate Crash Modeling for Motor Vehicle and Non-motorized Models at the Macroscopic Level. Accident Analysis and Prevention. Vol. 78. pp. 146-154.
27. Mannering, F. and Bhat, C., 2014. Analytic Methods in Accident Research: Methodological Frontier and Future Directions. Analytic Methods in Accident Research 1. pp. 1-22.
28. Rue, H. and Martino, S., 2007. Approximate Bayesian Inference for Hierarchical Gaussian Markov Random Field Models. Journal of Statistical Planning and Inference. Vol. 137, No. 10. pp. 3177-3192.

29. Aitchison, J. and Ho, C. H., 1989. The Multivariate Poisson Lognormal Distribution. Biometrika, Vol. 76, No. 4. pp. 643-653.

30. The R Project for Statistical Computing. https://www.r-project.org/

31. Qin, X., Wang, K. and Cutler, C., 2013. Logistic Regression Models of the Safety of Large Trucks. Transportation Research Record 2392. pp. 1-10.

32. Wang, K., Yasmin, S., Konduri, K., Eluru, N and Ivan, J., 2015. Copula-based Joint Model of Injury Severity and Vehicle Damage in Two-vehicle Crashes. Transportation Research Record 2514. pp. 158-166.

33. Connecticut Crash Data Repository (CTCDR). http://www.ctcrash.uconn.edu/

4.9 APPENDIX: PARAMETER ESTIMATION AND EXPLANATION

In the MVPLN model, the mean of the Poisson distribution for crash counts by crash type or crash severity is shown in Equation (4-2) as:

$$ln(\lambda_{ji}) = \Omega + \boldsymbol{z}'_{ji}\boldsymbol{\beta}_j + \gamma_{ji}$$

and we assume the error term $\boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i} \dots, \gamma_{Ji})'$ follows a *J*-dimensional normal distribution, which is expressed as:

$$\boldsymbol{\gamma}_i \sim Normal(\mathbf{0}, \boldsymbol{\Sigma}),\ \boldsymbol{\Sigma}=(\sigma_{rs})_{1\le r\le s\le J} \text{ and } \sigma_{rs} = \rho_{rs}\sqrt{\sigma_{ss} * \sigma_{rr}}$$

In terms of the equations above, the unknown parameters to be estimated in the MVPLN model are: $\boldsymbol{\beta}_j{}'\boldsymbol{s}$ which represent the posterior means of coefficients for all independent variables, σ_{ss} and σ_{rr} which represent the diagonal elements of variance-covariance matrix $\boldsymbol{\Sigma}$ for two crash type or severity (*s* and *r*) counts, and ρ_{rs} which represents the posterior mean of correlation coefficient between the r^{th} and s^{th} crash type or severity counts. Compared with the MVPLN model, the parameter estimates for the Univariate Negative Binomial (NB) model represent the coefficients estimated using the MLE approach with their standard deviations. The parameter

estimates for the Univariate Poisson Lognormal (UPLN) are similar to the MVPLN model, which are the posterior means with posterior standard deviations. However, the covariance term in the MVPLN model is ignored in the UPLN model. Since the error term in the MVPLN model includes a covariance part between two crash type or severity counts, the MVPLN model might explain extra variation of the data and lead to more accurate variance estimation for model parameters. The univariate models might declare insignificant predictors as significant and vice versa, by estimating incorrect variance.

The marginal distribution of the observed crash counts is shown in Equation (4-8) as:

$$g\left(\boldsymbol{y}_i\middle|\boldsymbol{z'}_{ji}\boldsymbol{\beta}_j,\boldsymbol{\Sigma}\right)=\int\dots\int f_{Normal,J}(\boldsymbol{\gamma}_i|\mathbf{0},\boldsymbol{\Sigma})\prod_{j=1}^{J}f_{Poisson}(y_{ij}|\gamma_{ji},\boldsymbol{z'}_{ji}\boldsymbol{\beta}_j)d\boldsymbol{\gamma}_i$$

There is no closed algebraic solution to the *J*-dimensional integral (20). We applied the model estimation through an approximate Bayesian framework. We firstly assume a Normal prior distribution for $\boldsymbol{\beta}_j{}'\boldsymbol{s}$ in Equation and a Wishart prior distribution for $\boldsymbol{\Sigma^{-1}}$ in Equation, then the posterior distribution of the parameters are estimated under the MVPLN regression model using the Bayesian inference. As described earlier, the Markov Chain Monte Carlo (MCMC) simulation approach is usually used to carry out fully Bayesian inference on the model coefficients. The MCMC uses the Gibbs sampler and Metropolis-Hastings (M-H) approach to carry out the Bayesian inference, which is extremely computationally challenging and time consuming especially for a large sample size and a dependent variable with many categories. To address this issue, we applied the INLA approach proposed by Rue *et al.* (2009) and Serhiyenko *et al.* (2016) to carry out the Bayesian inference. When approximating the posterior distributions of parameters, the INLA approach does not rely on the MCMC which can significantly reduce the model running time. Below is one selected example of using R-INLA package to estimate the MVPLN model for crashes by crash type on rural two-lane 3ST intersections, and the data

formulation used to estimate the model. The details of the R-INLA package are available at http://www.r-inla.org/ and Serhiyenko *et al.* (2016).

```
######################################################################
##############                                          ##############
##############    Multivariate modeling of 3ST Intersections ##############
##############                                          ##############
######################################################################

n=dim(mdata)[1]
N=4*n

## join crash counts together ##

y=c(mdata$SDC,mdata$IDC,mdata$ODC,mdata$SVC)

## create variables for multivariate analysis ##

LogOff=c(rep(mdata$LN_OFFSET,4))
Segment=rep(1:n,4)

LogAADT_MAJOR.SDC=c(mdata$LN_AADT_MAJOR,rep(NA,3*n))
LogAADT_MAJOR.IDC=c(rep(NA,n),mdata$LN_AADT_MAJOR,rep(NA,2*n))
LogAADT_MAJOR.ODC=c(rep(NA,2*n),mdata$LN_AADT_MAJOR,rep(NA,n))
LogAADT_MAJOR.SVC=c(rep(NA,3*n),mdata$LN_AADT_MAJOR)

LogAADT_MINOR.SDC=c(mdata$LN_AADT_MINOR,rep(NA,3*n))
LogAADT_MINOR.IDC=c(rep(NA,n),mdata$LN_AADT_MINOR,rep(NA,2*n))
LogAADT_MINOR.ODC=c(rep(NA,2*n),mdata$LN_AADT_MINOR,rep(NA,n))
LogAADT_MINOR.SVC=c(rep(NA,3*n),mdata$LN_AADT_MINOR)

LIGHTING.SDC=c(mdata$LIGHTING,rep(NA,3*n))
LIGHTING.IDC=c(rep(NA,n),mdata$LIGHTING,rep(NA,2*n))
LIGHTING.ODC=c(rep(NA,2*n),mdata$LIGHTING,rep(NA,n))
LIGHTING.SVC=c(rep(NA,3*n),mdata$LIGHTING)

APPROACH_LEFTTURN.SDC=c(mdata$APPROACH_LEFTTURN,rep(NA,3*n))
APPROACH_LEFTTURN.IDC=c(rep(NA,n),mdata$APPROACH_LEFTTURN,rep(NA,2*n))
APPROACH_LEFTTURN.ODC=c(rep(NA,2*n),mdata$APPROACH_LEFTTURN,rep(NA,n))
APPROACH_LEFTTURN.SVC=c(rep(NA,3*n),mdata$APPROACH_LEFTTURN)

APPROACH_RIGHTTURN.SDC=c(mdata$APPROACH_RIGHTTURN,rep(NA,3*n))
APPROACH_RIGHTTURN.IDC=c(rep(NA,n),mdata$APPROACH_RIGHTTURN,rep(NA,2*n))
APPROACH_RIGHTTURN.ODC=c(rep(NA,2*n),mdata$APPROACH_RIGHTTURN,rep(NA,n))
APPROACH_RIGHTTURN.SVC=c(rep(NA,3*n),mdata$APPROACH_RIGHTTURN)


## create intercepts ##

b1=c(rep(1,n),rep(NA,3*n))
b2=c(rep(NA,n),rep(1,n),rep(NA,2*n))
b3=c(rep(NA,2*n),rep(1,n),rep(NA,n))
b4=c(rep(NA,3*n),rep(1,n))
```

create crash index

```
i = 1:N
```

create dataset

```
dataall=data.frame(y,b1,b2,b3,b4,LogOff,LogAADT_MAJOR.SDC,
LogAADT_MAJOR.IDC,LogAADT_MAJOR.ODC, LogAADT_MAJOR.SVC,

LogAADT_MINOR.SDC,LogAADT_MINOR.IDC,LogAADT_MINOR.ODC,LogAADT_MINOR.SVC,
LIGHTING.SDC,LIGHTING.IDC,LIGHTING.ODC,LIGHTING.SVC,
APPROACH_LEFTTURN.SDC,APPROACH_LEFTTURN.IDC,
APPROACH_LEFTTURN.ODC,APPROACH_LEFTTURN.SVC,
APPROACH_RIGHTTURN.SDC,APPROACH_RIGHTTURN.IDC,
APPROACH_RIGHTTURN.ODC,APPROACH_RIGHTTURN.SVC)

formula = y ~ -
1+offset(LogOff)+b1+b2+b3+b4+LogAADT_MAJOR.SDC+LogAADT_MAJOR.IDC+LogAADT_MAJOR.ODC
+LogAADT_MAJOR.SVC+LogAADT_MINOR.SDC+LogAADT_MINOR.IDC+LogAADT_MINOR.ODC+Log
AADT_MINOR.SVC+LIGHTING.SDC+LIGHTING.IDC+LIGHTING.ODC+LIGHTING.SVC+APPROACH_LE
FTTURN.SDC+APPROACH_LEFTTURN.IDC+APPROACH_LEFTTURN.ODC+APPROACH_LEFTTURN.SV
C+APPROACH_RIGHTTURN.SDC+APPROACH_RIGHTTURN.IDC+APPROACH_RIGHTTURN.ODC+APPR
OACH_RIGHTTURN.SVC+f(i, model="iid4d", n=N)

MVPLN_TYPE = inla(formula, family="poisson",data = dataall, control.inla=list(h = 1e-5,tolerance = 1e-
3),quantiles=c(0.05,0.5,0.95),control.compute=list(dic=TRUE),control.predictor=list(compute=TRUE,link = 1))

summary(MVPLN_TYPE)
```

The data used to estimate the MVPLN model is formatted as below:

LN_AADT_MAJOR	LN_AADT_MINOR	LIGHTING	APPROACH_LEFTTURN	APPROACH_RIGHTTURN	SVC	SDC	IDC	ODC
8.14	4.02	0	0	0	0	0	0	0
8.14	6.01	0	0	0	1	0	0	1
8.73	7.52	0	1	1	5	1	1	0
8.99	8.36	1	1	1	2	3	3	0
8.25	6.50	0	1	1	4	0	0	0
8.60	7.79	1	1	1	1	1	2	3
8.33	7.27	1	1	1	1	0	0	1
……								

5 CONCLUSIONS AND CONTRIBUTIONS

In order to implement effective countermeasures to improve highway safety, it is necessary to determine appropriate methodologies for estimating both crash counts and crash severity, and identify the critical contributing factors to crashes for each facility type. Although the HSM provides methods for some roadway facilities, such as two lane rural highways, multilane rural highways, urban and suburban arterials, freeways and freeway ramp junctions, it has no guidance on predicting crashes for local road jurisdiction, where the traffic counts are generally not available. In addition, the method offered by the HSM estimating crashes in total might not be appropriate, as it doesn't account for the potential correlation between crashes among crash types and crash severities. This study explores sound statistical approaches for both crash count and crash severity analyses to address these issues, and addresses three objectives:

1. Modeling injury severity and vehicle damage simultaneously using a copula based model, and exploring the correlation between injury severity and vehicle damage across crashes.
2. Estimating TAZ level SPFs for both intersection and segments on roads under local jurisdiction in the State of Connecticut where the AADT is not available. The SPFs are estimated using socio-economic data and network topological data as a replacement for traffic count data, and the TAZs are categorized into different clusters based on the percentage of three land cover categories – high, medium and low intensities – and the population density (*i.e.* the number of population per km^2) to account for data heterogeneity.
3. Estimating crashes by both crash type and crash severity for 3ST intersections, 4ST intersections, 4SG intersections and roadway segments on rural two-lane highways using the MVPLN model, and exploring the possible correlations among crash type or crash severity counts.

The first study of this research demonstrates the use of six copula based models (including Gaussian, FGM, Frank, Clayton, Joe and Gumbel copula models) to explore the interrelationships between the two crash indicators: injury severity and vehicle damage, and also identify the nature of these correlations across different types of crashes. The study shows that the copula based models had a better goodness-of-fit than the independent model which indicates the existence of dependencies between injury severity and vehicle damage. Among the six copula based models, the Gaussian copula model had the best model performance with the lowest BIC value. The parameterized Gaussian copula structure shows that dependencies between injury severity and vehicle damage are positive for head-on, angle and sideswipe crashes, while the dependencies are negative for the crashes between two passenger cars, which indicates that the dependencies between injury severity and vehicle damage can vary across different crashes. The findings of this study show that when simultaneously estimating injury severity and vehicle damage, the correlations between these two indicators should be considered to get a more accurate model structure and coefficient estimates.

The second study of this research describes estimation of cluster based SPFs for local road intersections and segments in Connecticut using socio-economic and network topological data instead of traffic counts as exposure. The study indicates that the cluster-based SPFs outperform the aggregate SPFs using the entire dataset in crash prediction accuracy, and SPFs including total population, retail and non-retail employment and average household income are found to be the best in crash prediction performance. Finally, the cluster-based SPFs were applied and adjusted using the EB method to produce expected annual crash counts for all TAZs in the State of

Connecticut, and an ArcGIS visualization tool was developed using the clustered-based SPFs to help conduct roadway safety analyses.

The third study of this research describes estimation of crashes by both crash type and crash severity on rural two-lane highways, using the Multivariate Poisson Lognormal (MVPLN) model, and implementing approximate Bayesian inference via the Integrated Nested Laplace Approximation (INLA). The study shows that the INLA MVPLN model can significantly decreases the computational time compared with a fully Bayesian fitting of the MVPLN model using Markov Chain Monte Carlo (MCMC) methods. The coefficient estimates show that the standard errors of covariates in the MVPLN model are slightly lower than the other two univariate crash prediction models (*i.e.* Negative Binomial model and Univariate Poisson Lognormal model) when the covariates are statistically significant, and the crash counts by crash type and severity are significantly correlated. The model prediction comparisons illustrate that the MVPLN model outperforms the UPLN and NB model in prediction accuracy. The findings of this study indicate that when predicting crash counts by crash type and crash severity for rural two-lane highways, the MVPLN model should be considered to avoid estimation error and to account for the potential correlations among crash type counts and crash severity counts.

In summary, three major contributions are made in this research:

1. This research offers a more accurate model structure of predicting injury severity and vehicle damage, and can verify the different correlations between injury severity and vehicle damage across crashes, while also verifying the potential contributing factors affecting crash severity to shed light on developing cost-effective countermeasures or appropriate driver education to

mitigate or reduce the effects of crashes on drivers caused by these crash contributors. Specifically, considering vehicle damage - an objective indicator of crash severity might offer additional insight into the exploration of injury severity, which can be a biased and subjective indicator based on victim's responses, descriptions, and complaints after the crash. This research helps to enhance crash severity analysis, by compiling vehicle damage information in crash severity models, and exploring the effects of vehicle damage on injury severity suffered by occupants. Additionally, this research found that dependencies between injury severity and vehicle damage are negative for the crashes between two passenger cars. In other words, injury severity decreases when vehicle damage increases for crashes between two passenger cars, which indicates that the structural design of the vehicle can protect occupants from sustaining severe injuries by reducing or redirecting impact energies around occupants in a crash. This finding can help develop appropriate automotive features that reduce the possibility of more severe injuries sustained by occupants.

2. This research demonstrates an approach in developing TAZ level SPFs using socio-economic data and network topological data for local roads when the traffic volumes are not available, by clustering TAZs into different types to account for the data heterogeneity. The SPFs can be used to predict the average annual intersection and segment crashes in a TAZ in the context of HSM analyses, and might be used to help agencies evaluate alternative options for roadway network and economic development. In specific, the cluster-based SPFs can be applied as a planning tool to estimate the expected annual local road intersection and segment crashes under different development scenarios in the State of Connecticut, and the methods can be used to develop models for application elsewhere. Additionally, the estimated crash counts can be used in the developed ArcGIS visualization tool to help identify areas of cities

and towns with higher expected crash counts for local roads, and implement countermeasures to improve safety for city and town roads.

3. This research develops an approach to rigorously estimate crashes by crash type and crash severity in the context of current HSM analyses, and can identify the influence of roadway geometric characteristics on crash counts by crash type and crash severity, to help agencies assess roadway design alternatives, implement roadway safety facilities to reduce traffic crashes, especially the severe ones. One important finding of this research is that for rural two-lane segments, wider lane width is associated with increased opposite-direction crashes, but it is associated with decreased single-vehicle crashes. Roadway segments with higher speed limit are associated with decreased crashes for same-direction and opposite-direction crashes, but are associated with increased single-vehicle crashes. These different impacts of roadway geometric factors for different crash types or severities cannot be identified when modeling crashes in total, which suggests that estimating crashes by crash type or severity might be more helpful and accurate in identifying different crash contributing factors, compared with the original approach that estimating crashes in total.

Although the methodologies in this research are approved to be more reasonable in crash count or severity prediction, there are still some limitations that cannot be addressed in this research. The copula based model employs only two vehicle crashes for the analysis. The findings may not be directly transferable to crashes involving single vehicles or more than two vehicles. The cluster based TAZ SPFs developed for the State of Connecticut may be difficult to be transferred to other jurisdictions, as the TAZ level SPFs are highly dependent upon not only the clustering of the TAZs, but also the definitions of the TAZs themselves, as well as the character of land

development. Other relevant variables that were not available (*e.g.* trip distance and trip duration for a TAZ) may also affect roadway safety and SPF estimation. Additionally, the data sample for 4SG intersections in the third study is too small to compare the MVPLN model with the UPLN and NB models regarding the prediction performance. Therefore, to enhance the safety analysis accuracy, future research can focus on the new analyses stated as follows.

1. New analyses can focus on collecting crash severity data for not only two-vehicle crashes, but also single-vehicle and multi-vehicle crashes to estimate the copula based model, and identify if the crash contributing factors and dependencies between injury severity and vehicle damage for single-vehicle and multi-vehicle crashes are consistent with two-vehicle crashes, which helps verify whether different copula-based models should be considered in crash severity analyses for crashes involve different number of vehicles.
2. New analyses can focus on collecting crashes on local roads from the States other than Connecticut, and estimate cluster based SPFs to test the model transferability. It is also recommended to collect extra variables (*e.g.* trip distance and trip duration for a TAZ) that are not available in this research, and then estimate new SPFs to improve prediction accuracy. New analyses can also focus on collecting traffic volume information for local roads, and develop traditional SPFs using AADT as exposure to compare the results with SPFs using socio-economic data and network topological data.
3. New analyses can focus on collecting more rural two-lane 4SG intersections, and then estimate new MVPLN models for rural two-lane 4SG intersections to improve prediction accuracy. Future analyses can also target on estimating MVPLN models by crash type and severity for other roadway facility types, such as rural multilane highways, urban and suburban arterials, and freeways to identify the different impacts of roadway geometric

factors for different crash types or severities for these roadway facilities, and help enhance traffic safety analysis accuracy and implement safety improvement projects for each roadway facility type. Furthermore, although the MVPLN model can provide unbiased and more accurate variance estimates for parameters than the univariate models, the MVPLN model highly depends on the assumption that the crash type or severity counts are correlated, and usually be used when the crash type or crash severity counts are estimated simultaneously. It may not be necessary to use the MVPLN model when crash counts are estimated independently, and the MVPLN model is hard to be estimated and implemented due to the complex model structure. Future research can focus on developing a detailed model selection process to indicate under which assumptions (*i.e.* whether the crash counts are treated as independent variables or dependent variables) and crash estimation objectives (*i.e.* whether crash counts are estimated in total or crash counts are estimated by crash type and severity) the MVPLN model could be considered instead of univariate models, and vice versa. Future research can also focus on developing convenient programming tools (such as R, Python and GAUSS programming codes) and application tools (such as ArcGIS and TransCAD planning tools) to help users easily estimate the MVPLN model and implement the model in highway safety analyses.