Missouri S&T Library and Learning Resources

Scholars' Mine

Doctoral Dissertations | Student Theses and Dissertations

2014

Longitudinal analysis of crash frequency data

Mojtaba Ale Mohammadi

Follow this and additional works at: https://scholarsmine.mst.edu/doctoral_dissertations

 Part of the Civil Engineering Commons

Department: Civil, Architectural and Environmental Engineering

Recommended Citation

Ale Mohammadi, Mojtaba, "Longitudinal analysis of crash frequency data" (2014). *Doctoral Dissertations*. 2498.
https://scholarsmine.mst.edu/doctoral_dissertations/2498

This thesis is brought to you by Scholars' Mine, a service of the Missouri S&T Library and Learning Resources. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

www.manaraa.com

LONGITUDINAL ANALYSIS OF CRASH FREQUENCY DATA

by

MOJTABA ALE MOHAMMADI

A DISSERTATION

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

2014

Approved by
Dr. V.A. Samaranayake, Advisor
Dr. Ghulam Bham, Co-Advisor
Dr. Ronaldo Luna
Dr. Robert Paige
Dr. Abhijit Gosavi

www.manaraa.com

Copyright 2014
Mojtaba Ale Mohammadi
All Rights Reserved

www.manaraa.com

DEDICATION

I would like to dedicate this doctoral dissertation to my parents Nasser Ale Mohammadi and Maryam Rahimian, my two sisters Zakieh and Zahra, and my fiancée Mina Hadi whose continued love and support helped me much in completion of this process.

www.manaraa.com

PUBLICATION DISSERTATION OPTION

This dissertation has been prepared in the styles utilized by the Transportation Research Record: Journal of Transportation Research Board and Journal of Analytic Methods in Accident Research. Pages 5-21 are published in Transportation Research Record: Journal of the Transportation Research Board; Pages 22-59 will be published at the Analytic Methods in Accident Research; Pages 60-85 is submitted to the Transportation Research Record: Journal of Transportation Research Board. Appendices A and B and the Bibliography have been added for purposes normal to dissertation writing. Appendix C was also added last to incorporate the addressed comments of the committee members on the published papers.

www.manaraa.com

ABSTRACT

This study comprises mainly of three papers. First, a systematic evaluation of the effects of Missouri's Strategic Highway Safety Plan between 2004 and 2007 is presented. Negative binomial regression models were developed for the before-through-change conditions for the various collision types and crash severities. The models were used to predict the expected number of crashes assuming with and without the implementation of MSHSP. This procedure estimated significant reductions of 10% in crashes frequency and a 30% reduction for fatal crashes. Reductions in the number of different collision types were estimated to be 18-37%. The results suggest that the MSHSP was successful in decreasing fatalities.

Second, ten years (2002 - 2011) of Missouri Interstate highway crash data was utilized to develop a longitudinal negative binomial model using generalized estimating equation (GEE) procedure. This model incorporated the temporal correlations in crash frequency data was compared to the more traditional NB model and was found to be superior. The GEE model does not underestimate the variance in the coefficient estimates, and provides more accurate and less biased estimates. Furthermore, the autoregressive correlation structure used for the temporal correlation of the data was found to be an appropriate structure for longitudinal type of data used in this study.

Third, this study developed another longitudinal negative binomial model that takes into account the seasonal effects of crash causality factors using Missouri crash data. A GEE with autoregressive correlation structure was used again for model estimation. The results improve the understanding of seasonality and whether the magnitude and/or type of various effects are different according to climatic changes. It was found that the traffic volume has a higher effect in increasing the crash occurrence in spring and lower effect in winter, compared to fall season. The crash reducing effect of better pavements was found to be highest in spring season followed by summer and winter, compared to the fall season. The results suggest that winter season has the highest effect in increasing crash occurrences followed by summer and spring.

www.manaraa.com

ACKNOWLEDGMENTS

Although this dissertation is individual work, I could have never reached the heights and explored the depth without the help and support of many people.

Firstly, I would like to thank my Advisor Dr. Samaranayake for instilling in me the qualities of being a good engineer who can plan for an objective and organize tasks and times in advance to achieve the goals. Dr. Sam has opened my eyes to the aspects of research which were not so vivid to me previously. He made me better understand that discussion of the results of an experiment is what makes it meaningful; always good to look at and argue the results and not simply see them as the possible outcomes.

I would like to thank my Co-advisor as well Dr. Bham who made me develop the required skills for being always hopeful in times of denial and disapproval by the higher ranked authorities, for changing my point of view toward people who demand without the appreciation of encouragement and motivation. He taught me to be self-inspired and love what I do.

I thank my committee members Dr. Ronaldo Luna, Dr. Robert Paige, and Dr. Abhijit Gosavi for their support. I also thank Karen White the graduate secretary of the civil engineering department for her many helpful tips.

I would like to thank Dr. Ezra Hauer for his father like and professional advice during a three-day workshop on safety performance functions. I also would like to thank Dr. Dominique Lord, Dr. Fred Mannering, Dr. Mohamed Abdel-Aty, and Dr. Venkataraman Shankar for their in-depth research on safety analysis which were very helpful to me.

I would like to thank Amirhossein Rafati and Alireza Toghraei for their understanding and not only being my roommates but also good friends as well. I would also like to present my thankfulness to the good friends whose support drove me forward everyday: Amin Assareh, Nima Lotfi, Hesam Zomorodi, Maryam Kazemzadeh, Kousha Marashi, Hossein Sepahvand, Hassan Golpour, Mansoureh Fazel, and Farzin Ferdowsi and other loving friends.

www.manaraa.com

TABLE OF CONTENTS

Page

DEDICATION iii

PUBLICATION DISSERTATION OPTION iv

ABSTRACT v

ACKNOWLEDGMENTS vi

LIST OF ILLUSTRATIONS ix

LIST OF TABLES x

SECTION

1. INTRODUCTION 1

PAPER

I. SAFETY EFFECT OF MISSOURI'S STRATEGIC HIGHWAY SAFETY PLAN - MISSOURI'S BLUEPRINT FOR SAFER ROADWAYS 5

ABSTRACT 5

1. INTRODUCTION 6

2. BACKGROUND 7

3. METHODOLOGY 8

4. DATA ANALYZED 11

5. RESULTS 13

6. CONCLUSIONS AND RECOMMENDATIONS 18

7. REFERENCES 19

II. CRASH FREQUENCY MODELING USING NEGATIVE BINOMIAL MODELS: AN APPLICATION OF GENERALIZED ESTIMATING EQUATION TO LONGITUDINAL DATA 22

ABSTRACT 22

1. INTRODUCTION 23

2. METHODOLOGY 26

3. CRASH DATA 31

3.1. Data Description 31

3.2. Multicollinearity 34

3.3. Sample Size 35

www.manaraa.com

3.4. Confounding Effects and Variable Specification 36
4. RESULTS AND DISCUSSION 40
4.1. Model estimates and comparisons 40
4.2. Validation of correlation structure 49
5. CONCLUSIONS 51
6. ACKNOWLEDGEMENTS 52
7. REFERENCES 53
III. SEASONAL EFFECTS OF CRASH CONTRIBUTING FACTORS ON HIGHWAY SAFETY 60
ABSTRACT 60
1. INTRODUCTION 60
2. METHODOLOGY 62
3. CRASH DATA AND MODEL VARIABLES 64
4. RESULTS AND DISCUSSION 69
5. CONCLUSIONS AND RECOMMENDATIONS 79
7. REFERENCES 81
SECTION
2. CONCLUSIONS 86
APPENDICES
A. MATLAB ALGORITHM FOR READING THE CRASH DATA BASE, ROAD INVENTORY DATA BASE, ASSIGNING SEGMENT IDENTIFICATIONS, AGGREGATING YEARLY, MONTHLY, AND SEASONAL CRASH FREQUENCY 91
B. SAS CODES FOR MODELING CRASH FREQUENCY. 100
C. DETAILS OF THE EXAMINATION FOR CONFOUNDING AND SUFFICIENCY OF OBSERVATIONS. 115
D. PAPER CORRECTIONS ADDENDUM. 132
BIBLIOGRAPHY 135
VITA 143

www.manaraa.com

LIST OF ILLUSTRATIONS

Figure Page

Paper I

1 Graphical comparison of the effect of safety improvements on the crash models developed. 16

2 Percent reduction in the expected number of crashes predicted in 2008 as a result of safety improvement strategies. 18

Paper II

1 Total number of crashes on a selected few of the interstate highways of Missouri with most variation 33

2 Estimates of the interaction terms between number of lanes, and speed limit in urban areas 43

3 Comparison of the models' standard errors using generalized estimating equations and maximum likelihood estimation methods 47

4 Comparison of the models' $\chi 2$-values using generalized estimating equations and maximum likelihood estimation methods 48

5 Cumulative residuals plot for LnAADT for the negative binomial models estimated using the methods of generalized estimating equation and maximum likelihood estimation 49

Paper III

1 Estimates of the interaction terms between number of lanes, and speed limit in urban areas 74

2 Estimates of the significant interaction terms between number of lanes, and speed limit in urban areas during winter season 77

www.manaraa.com

LIST OF TABLES

Table Page

Paper I

1 Descriptive statistics of segment properties of interstate highways in Missouri 12

2 Parameter estimates and their standard errors for the different negative binomial models 14

3 Comparison of the predicted crash count properties for 2008 with/without safety improvements 17

Paper II

1 Correlation values for the autoregressive Type 1 and exchangeable structure 29

2 Descriptive statistics of segment properties of Missouri interstates 32

3 Pearson correlation coefficients and collinearity diagnostics 34

4 Number and percentage of observations within area types, by number of lanes and speed limit 36

5 List of the dummy variables considered for the analysis 38

6 List of the continuous and classification variables considered for the analysis 39

7 NB model estimates 42

8 Analysis of statistical significance of the effect of change in the number of lanes and speed limit on crash frequency 45

9 Statistical significance of the effect of change in the number of lanes and speed limit in the model estimated by the method of maximum likelihood estimation 46

10 Comparison of relatively smaller $\chi 2$-values 48

11 Negative binomial model estimates using generalized estimating equations for three different analysis periods 50

Paper III

1 Definition of the continuous and dummy variables considered for analysis 67

2 Definition of the interaction variables considered for analysis 68

3 Negative binomial model parameter estimates 70

4 Overall amount and statistical significance of the effect of change in the number of lanes and speed limit on crash frequency in urban areas of Missouri 75

5 Amount and statistical significance of the effect of change in the number of lanes and speed limit on crash frequency in urban areas during the winter season 78

www.manaraa.com

SECTION

1. INTRODUCTION

Traffic safety in transportation networks is one of the main priorities for many government agencies, private organizations and the society as a whole. This is mainly due to the significant monetary and non-monetary costs associated with crashes (Elvik, 2000). According to the National Highway Traffic Safety Administration, 5,505,000 traffic crashes occurred in 2009 on the US highways in which 33,808 people died and 2,217,000 people were injured (NHTSA, 2009). Peden et al. (2004) found that the trend in road related injuries are expected to increase from ranked ninth in 1990 to the third largest contributor to the global burden of disease and injury in 2020. This immense loss to society resulting from motor vehicle crashes warrants careful crash evaluation and safety analysis to accurately identify crash contributing factors and countermeasures. HSM (2010) regards crash frequency as a fundamental indicator of "safety" in terms of evaluation and estimation.

Crash analysis research in general has focused on the estimation of traditional crash prediction models such as negative binomial (NB) and Poisson regression models and their generalized forms due to their relatively good fit to the data (Shankar et al., 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Savolainen and Tarko, 2005; Mojtaba Ale Mohammadi et al., 2014a). These crash prediction models have also been used for crash evaluation purposes. HSM (2010) refers to the term "crash evaluation" as the process of determining the effectiveness of a particular treatment after its implementation. Many studies have been conducted to investigate the effect of improvement programs on facilities such as rail-highway grade crossings (Hauer and Persaud, 1987), highway segments (Zegeer and Deacon, 1987; Squires and Parsonson, 1989; Knuiman et al., 1993), and intersections (Poch and Mannering, 1996; Datta et al., 2000). One of the issues that has been raised regarding the use of these traditional models on "crash evaluation" is the statistical phenomenon of regression to the mean that occurs when the same unit of observations is repeatedly measured over time (Barnett et al.,

www.manaraa.com

2005). This phenomenon may result in biased estimates in any such investigations and mask the real effectiveness of any countermeasure which in turn clouds the judgment of the evaluators and results in unwise decisions. Empirical Bayes (EB) method has also been used for before-after studies to evaluate the effect of countermeasures on safety, which properly accounts for the regression to the mean while normalizing for differences in traffic volume and other factors between the before and after periods (Hauer, 1997; Persaud et al., 2004; Guo et al., 2010b; Shively et al., 2010; Yu et al., 2013b). But EB method is a relatively sophisticated method that requires extensive data and considerable training and experience (Persaud and Lyon, 2007).

This study presents a simple new approach to addresses the problems mentioned above. A traditional negative binomial regression model was developed using the introduced method to examine the effect of implementation of the Missouri Strategic Highway Safety Plan (MSHSP). The MSHSP data was chosen as it provides an excellent situation of safety improvement intervention on the highways of Missouri. In addition it and evaluate the effects of MSHSP on the crash frequency of various collision types and severity levels. The negative binomial regression models were developed to account for the before-through-change conditions using a continuous variable that is set to zero for pre-implementation years and gradually increases over the implementation years to reach a plateau at the conclusion of the plans.

In the second section of this study, a (longitudinal) negative binomial model was developed using ten years of data (2002-2011). Lord and Persaud (2000) suggest that more years of data adds up to the reliability of the model estimates by reducing the standard errors in the prediction models mentioned above; However, when many years of data is considered the serial correlation in the repeated observations violates the independence assumptions on unobserved error terms in traditional Poisson and/or NB crash frequency models. This violation creates biased and inefficient models by underestimating the standard errors. Researchers have tried to use different techniques to account for these temporal correlations between the repeated crash frequencies observed for a highway segment over the years. Examples of the utilized methodologies can be found in Maher and Summersgill (1996) using an iterative solution based on the method of "constructed variables" presented by McCullagh and Nelder (1989), in Ulfarsson and

www.manaraa.com

Shankar (2003) using negative multinomial (NM) models, in Dong, Richards, et al. (2014) using multivariate random parameter models, and in Venkataraman et al. (2014) using random parameter negative binomial models. These methodologies, however, have shown to be not practically applicable for different situations. For example, the analyst need to know the extent and type of correlation prior to the analysis that is not always known (Lord and Persaud, 2000), or the estimation methodology for multivariate random parameter models – the full Bayesian method – is complex and requires training and practice. The implementation and transferability of the method is also a challenge. Wang and Abdel-Aty (2006) used generalized estimating equations (GEE) technique to account for these correlations in a frequency model for rear-end crashes at signalized intersections. This technique has the potential of addressing the issue of serial correlations in the repeated observations, producing reasonably accurate standard errors and efficient parameter estimates (Méndez et al., 2010; Peng et al., 2012; Giuffrè et al., 2013; Stavrinos et al., 2013). Liang and Zeger (1986) were the first to use this technique to model repeated observations and showed that the GEE method is robust to misspecification of the correlation structure but Giuffrè et al. (2007) and Ballinger (2004) demonstrated that utilizing the true data correlation structure in safety modeling results in higher estimation precision. In spite of all this research on the effects of temporal correlations in crash data, consequences arising from the omission of the serial correlation are still not completely understood.

The longitudinal negative binomial model developed in the second part of this study presents an application of the GEE method to model several years of crash frequency data in Missouri. This analysis first determines the temporal correlation structure in the data, proceeds with the analysis, and finally validates the correlation structure used in the analysis as an appropriate structure in this type of data. During the analysis, several data-related obstacles had to be addressed including the multicollinearity, sufficiency of the within-cluster observations, and the confounding effects. Interaction of the major crash contributing factors with the area type was also examined to evaluate whether crash causes behave differently from rural to urban areas. The results of this model were then compared with a traditional NB model using the chi-

www.manaraa.com

square values of the estimated model parameters and the cumulative residual (CURE) plots. Details of this part of the study are presented in the section "Paper II".

The results of the second section provide a better understanding of the true factors that affect the occurrence of crashes. The third part of this study is also involved in improvements of the crash evaluation model. Crashes are usually caused by several factors related to drivers' behavior, vehicles, highway design, and environmental conditions. Geographic location and the climatic environment, particularly seasonal weather can be a major factor that contributes to the occurrence of crashes (Garber and Hoel, 2008b). There are few studies in the crash evaluation realm dealing with the seasonal effects of crashes, but to the best knowledge of the author, no in-depth analysis of the seasonality of crash causes has been conducted. Some examples of the previous studies on the seasonality effects include the works of Carson and Mannering (2001), Hilton et al. (2011), Ahmed et al. (2011), Yu et al. (2013a), and Yang et al. (2013) that have shown that with a better understanding of the crash causes over different times of the year, policy-makers can improve the safety of specific roadway segments according to the seasonal weather patterns and that different traffic management strategies should be designed based on seasons.

The objective of the analysis in the third paper is to further investigate the seasonal effects on crash causality factors by developing a longitudinal negative binomial model using ten years of crash data on six main interstate highways of Missouri. This analysis uses generalized estimating equation (GEE) technique to develop the model. The interaction of the main variables with the seasonal indicators were examined in the model to gain a better understanding of the change in the effect of crash causes over different seasons in a year. The effects of interventions made by the Missouri Strategic Highway Safety Plan (MSHSP) over the years 2005-2011 is also investigated. The detailed results of this analysis (presented in the section "Paper III") can help in developing policies regarding highway safety countermeasures with insight on the effects of seasonal changes on roadway fatality factors.

www.manaraa.com

PAPER

I. SAFETY EFFECT OF MISSOURI'S STRATEGIC HIGHWAY SAFETY PLAN - MISSOURI'S BLUEPRINT FOR SAFER ROADWAYS

ABSTRACT

This study systematically evaluates the changes in motor vehicle crashes that occurred on the Missouri interstate highway system following the implementation of Missouri's Strategic Highway Safety Plan (MSHSP) between 2004 and 2007. The MSHSP implemented crash injury reduction strategies in enforcement, education, engineering, and public policy. Empirical Bayesian methods are commonly used to evaluate the effects of any change in safety as a result of countermeasures. This study presents a simple new approach to evaluating the effects of Missouri's safety plans on roadway crashes. For crash data associated with traffic and roadway characteristics, negative binomial regression models were developed for the before-through-change conditions using a variable that is set to zero for pre-implementation years and gradually increases over the implementation years to reach a plateau at the conclusion of the safety plans. The models developed for the various collision types and crash severities were used to estimate the expected number of crashes at roadway segments in 2008, assuming with and without the implementation of MSHSP. This procedure estimated significant reductions of 10% in the overall number of crashes and a 30% reduction for fatal crashes. Reductions in the number of different collision types were estimated to be 18-37%. The theoretical results indicate that the MSHSP was a successful policy in reducing the number of crashes and decreasing fatalities by reducing the most severe collision types like head-on crashes. The results are also consistent with many international studies and suggest that the safety strategic plans should be promoted as an effective treatment for highways.

Keywords: negative binomial, before-after study, Missouri blueprint, strategic highway safety plan, MSHSP

www.manaraa.com

1. INTRODUCTION

In 2004, a partnership of Missouri safety advocates, including law enforcement agencies, health care providers, government agencies, and others formed the Missouri Coalition for Roadway Safety (MCRS). This group worked with regional safety coalitions to implement the first strategic highway safety plan, titled Missouri's Blueprint for Safer Roadways. The potentially life-saving and injury reduction strategies in Missouri's Blueprint were crucial in the areas of education, enforcement, engineering, and public policy. Some of these strategies included the increase in public education and information on traffic safety, expanding roadway shoulders, installation of centerline and shoulder rumble strips, and roadway visibility features such as pavement markings, signs, lighting, etc., removing fixed objects along roadside right of way, and improving curve recognition through the use of signs, markings, and pavement treatments.

The primary emphasis area of the program aimed to reduce the number and severity of serious crash types with a specific focus on run-off-road crashes, crashes involving horizontal curves, head-on crashes, collisions with trees or poles, and intersection crashes (*1*). The long-range goal of the program was to reach 1000 or fewer fatalities by 2008 which was achieved a year early, when the total number of fatalities was reduced to 992 in 2007. Between 2005 and 2007, the death rate per 100 million vehicle miles of travel dropped from 1.8 to 1.4 and 21% fewer lives were lost on Missouri highways (*2*). These safety improvements resulted from the implementation of the MSHSP (*1*, *2*). The present study theoretically examines the effect of implementation of the Missouri's Blueprint for Safer Roadways on the nature and magnitude of crash frequency of various collision types and their severity. The next section presents a review of the previous studies in the literature of highway safety. The paper then describes the approach used in this study along with an introduction to the data set used. The results of the study and the conclusions follow in the next sections.

www.manaraa.com

2. BACKGROUND

Highway safety analysts use regression models for purposes such as establishing relationships between motor vehicle crashes and incorporating factors such as traffic and geometric characteristics of the roadway, predicting values or screening variables (*3*). Lord and Mannering (*4*) have documented a considerable amount of research work devoted to the development and application of new and innovative models for analyzing count data. According to Zou et al. (*5*), due to the over-dispersion in crash data, the negative binomial (NB) model is probably the most frequently used statistical model in various types of highway safety studies for developing crash prediction models. Shankar et al. (*6*) conducted a negative binomial multivariate analysis of roadway geometrics and weather-related effects. Their work presents a basis for a comprehensive before-and-after analysis of the effectiveness of safety improvements.

Developing quantitative relations to relate various safety improvement plans to crash rates and severities provides the information required to choose between the cost and the benefit of better transportation networks, and also helps in prioritizing the safety improvement projects. Many studies have been conducted in the past decades investigating the effect of improvement programs on facilities such as rail-highway grade crossings (*7*), highway segments (*8-10*), and intersections (*11*, *12*).

Researchers have also used the Empirical Bayes (EB) method (*13*) for conducting observational before-after studies to evaluate the effect of engineering countermeasures on safety. This procedure is often used to properly account for the regression to the mean while normalizing for differences in traffic volume and other factors between the before and after periods. Persaud et al. (*14*) used the EB procedure to examine the reduction of opposing direction crashes after installation of rumble strips along the centerlines of undivided rural two-lane roads. Bayesian inference methods have also been used in many recent studies to predict crash occurrences (*15*, *16*). Miaou et al. (*17*) and Ahmed et al. (*18*) employed the Hierarchical Bayes model to estimate traffic crashes. Shively et al. (*19*) employed a Bayesian nonparametric estimation procedure in their study. Huang and Abdel-Aty (*20*) also proposed a hierarchical structure to deal with multilevel traffic safety data. Persaud and Lyon (*21*) conducted extensive research on the EB methodology and its

www.manaraa.com

statistical application in before-after studies. According to them, there is a need to evaluate the safety effect of roadway improvements that may impact crash frequency, and the EB methodology produces valid results that are substantially different than those produced by more traditional methods. What requires exploration is whether or not it is worth the effort of using a sophisticated methodology such as the EB method in which (a) the relative complexity of the methodology requires analysts with considerable training and experience, and (b) the data needs can be extensive (*21*).

The more conventional alternatives to the EB method, involving a simple before–after comparison of crash counts or rates, with or without a comparison or control group, are appealing in that they are relatively easy to apply. These alternative methods, however, are loaded with challenges (*21*): the comparison group needs to be similar to the treatment group in all of the possible factors that could influence safety, and the assumption that the comparison group is unaffected by the treatment is difficult to test and can be unreasonable in some situations.

This study presents a simple new approach to evaluate the effects of MSHSP on Missouri Interstate highway crashes. Using six years of data, including the safety program implementation years (2005 through 2007), negative binomial crash frequency models were developed for predicting the crash frequency for 2008. The prediction models are developed in a way that will address the regression to the mean concern that prevails in such models. The predicted crash frequency with and without the improvements was compared statistically to determine the effect of MSHSP. The models represent a mix of urban and rural environments and were developed for various collision types and crash severities to investigate the safety improvements by estimating the expected number of crashes under different scenarios.

3. METHODOLOGY

The safety of an improved segment of the roadway in general should be estimated by mixing information of causal factors such as traffic flow, type of traffic control devices, geometric properties, etc. (*7*). The objective of this study is to develop statistical models of the crash frequency for all the interstate highways of Missouri. This study

www.manaraa.com

estimates six different crash frequency models that will predict (*1*) total crash frequency (all crash types), (2) head-on crash frequency, (3) rear-end crash frequency, (4) sideswipe-same direction crash frequency, (5) sideswipe-opposite direction crash frequency, and (6) angle crash frequency. Additionally, two separate models are developed for the only fatal and only non-fatal crashes. The dependent variable in all models is the crash count with a discrete non-negative integer nature, and Poisson regression is the first natural choice for modeling such data (22-25); however, a major limitation of the Poisson model is that it constrains the variance of dependent variables to be equal to its mean. When the variance of the data is not equal to the mean (which is usually the case in most of the crash frequency data), the variance of the model coefficients tend to be underestimated, which results in biased estimates. Negative binomial models have been extensively used in literature to overcome this limitation by relaxing the condition of 'variance = mean' in standard Poisson models (5).

If the length of segment 'i' (L_i) and the crash observation time interval for segment 'i' (t_i) for various segments are different, the observed number of crashes on the segment 'i' is proportional to the L_i and t_i. Length and duration of the observation are commonly called to be offset variables as their coefficients are restricted to be one and not estimated (26). In this study, since all the segments are measured over one year, the only offset variable used was the segment length. To describe the formulation of the negative binomial model, the Poisson model for crash counts is first reviewed; according to the Poisson distribution the probability of 'n' crashes occurring on segment 'i' during time period 'j' is:

$$P(n_{ij}) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{n_{ij}}}{n_{ij}!} \quad (1)$$

Where λ_{ij} is the expected number of crashes on segment 'i' during time interval 'j'. Given the vector of incorporating factors, λ_{ij} can be estimated by the equation:

$$\ln \lambda_{ij} = X_{ij}\beta \quad (2)$$

Where 'X' is a vector of affecting variables and 'β' is a vector of estimable coefficients.

www.manaraa.com

An additional stochastic component 'ε' is introduced to the link function by assuming 'e^{ε}' Gamma distributed (with mean 'μ' and variance 'α') resulting in the Poisson-Gamma model (also called the negative binomial model, NB) (*6*, *24*, *27*, *28*):

$$\ln \lambda_{ij} = X_{ij}\beta + \varepsilon_{ij} \tag{3}$$

An additional parameter 'α' allows the variance to differ from the mean and will result in the following mean-variance relationship:

$$var(n_{ij}) = E[n_{ij}]\left[1 + \alpha E[n_{ij}]\right] = \mu_i(1 + \alpha\mu_i) \tag{4}$$

If 'α' is equal to zero, the negative binomial reduces to Poisson, and if it is significantly different from zero, the data is either over-dispersed or under-dispersed. Using the Poisson distribution for crash count modeling, the probability of n crashes occurring on segment 'i' during time period 'j' is:

$$P(n_{ij}) = \frac{\Gamma(\theta + n_{ij})}{\Gamma(\theta)\mathrm{n_{ij}}!}\left(\frac{\theta}{\theta + \lambda_{ij}}\right)^{\theta}\left(\frac{\lambda_{ij}}{\theta + \lambda_{ij}}\right)^{n_{ij}} \tag{5}$$

Where $\theta = 1/\alpha$, and $\Gamma(.)$ is a value of gamma function. λ_{ij} can be estimated using the maximum likelihood estimation (MLE) procedure. The likelihood function for the negative binomial model is:

$$L(\lambda_{ij}) = \prod_{i=1}^{N}\prod_{j=1}^{T}\frac{\Gamma(\theta + n_{ij})}{\Gamma(\theta)\mathrm{n_{ij}}!}\left[\frac{\theta}{\theta + \lambda_{ij}}\right]^{\theta}\left[\frac{\lambda_{ij}}{\theta + \lambda_{ij}}\right]^{n_{ij}} \tag{6}$$

Where 'T' is the last time interval of the crash count data and 'N' is the number of roadway segments. Maximizing this function results in the estimation of 'β' and 'α' (in equations 2 and 3). Using a variable that is set to zero for pre-implementation years and gradually increases over the implementation years (2005 through 2007) to reach a plateau of one at the conclusion of the safety plans and the crash data associated with traffic and roadway characteristics, negative binomial regression models were developed for the before-through-change conditions.

www.manaraa.com

The reduction in the crash frequency after the implementation of the safety plans relative to the frequency values prior to this implementation could be attributed to the simple phenomenon of regression to the mean. If the reduction in the crash frequencies was detected using a model that uses before and after values, then associating this reduction with the implementation of the safety measures may be misleading. Our approach, however, did not merely look at before and after figures or model the change using a dummy variable, but instead utilized a continuous variable named "transition" in the NB model of the analysis to account for the plan implementation through the years. This variable was assigned the value of zero prior to the commencement of the improvements and gradually increased from zero to one, exactly over the implementation period in such a way that its value coincided, approximately, with the proportion of safety features that were completed at a given time. For the years after the completion of the improvements, this variable was kept constant at 1.0, suggesting 100% implementation. The plan included actions such as widening roadway shoulders, installation of centerline and shoulder rumble strips, etc. This study is an attempt to statistically examine the effects of the MSHSP implementation. The transition variable turned out to be highly significant with a negative sign for its coefficient estimate, indicating a close correlation between the reduction in crash frequency and the rate of completion of the safety features. Hence, the likelihood that this reduction reflects a regression to the mean is very low.

4. DATA ANALYZED

The Missouri Department of Transportation (MoDOT) portal of safety investigation provided access to the crash data base for all the recorded years of data. The crash data consists of all severity types of motor-vehicle crashes (fatal, disabling injury, minor injury, and property damage only crashes) at 17 interstate highways in the state of Missouri from 2002 to 2007. Some of the major characteristics of the highways are presented in Table 1. These highways, with an overall length of about 1200 miles, were classified as divided highways located either in urban or rural areas (65% in rural areas and 35% in urban areas). The total number of crashes in the data set analyzed was

www.manaraa.com

167,783 crashes, out of which 37% occurred in rural areas and 63% in urban areas. The rate of crash (number of crashes per mile and number of crashes per vehicle) for a segment in each year is shown in column 2 of the table along with the total number of crashes on all interstate highways presented in column 3.

Table 1. Descriptive statistics of segment properties of interstate highways in Missouri

Year	Crash/mile, Crash/1000 car* (min-max)	Total number of crashes	AADT (min-max)	Number of lanes (min-max)	PSR (min-max)	Percent commercial (min-max)
2002	0-241 , 0-4	18955	1985-101594	1-7	19.3-66.4	0.041-0.582**
2003	0-293 , 0-4	19581	1867-98485	1-7	17.4-37.4	0.041-0.406
2004	0-293 , 0-4	19343	1919-109420	1-6	18.9-37.3	0.046-0.582
2005	0-328 , 0-4	19101	1865-109573	2-6	24-39.6	0.045-0.582
2006	0-500 , 0-3	18922	1874-114753	1-6	23.4-37.5	0.049-0.582
2007	0-333 , 0-4	19308	1893-115901	1-6	22.9-37.6	0.049-0.622

* Minimum rate for all the segments during each year was zero

** This high value of truck percentage probably represents the night time at specific sections of the highways with low traffic

The explanatory variables used in this analysis are number of lanes, lane width (min. 10 ft to max. 18 ft), shoulder width (min. 3 ft to max. 12 ft), average annual daily traffic (AADT), speed limit, congestion index, pavement serviceability rate (PSR), and truck percentage. Other factors such as weather information, roadway conditions, and drivers' characteristics could not be aggregated for the entire state and yearly level for analysis. PSR is equal to two times the ride number plus the pavement condition index. Ride number is an index derived from controlled measurements of longitudinal profile in the wheel tracks and correlated with rideability of a pavement using a scale of 0 to 5, with 5 being perfect and 0 being impassable. Pavement condition index is a numerical rating of the pavement condition that ranges from 0 to 100 with 0 being the worst possible condition and 100 being the best possible condition. More information on the indices of ride number and pavement condition index can be found on the standards ASTM D6433-07 (*29*) and ASTM E1489-08 (*30*) respectively. The higher the value of PSR, the higher the pavement serviceability. Congestion index presents the level of congestion. It is

www.manaraa.com

calculated by incorporating the level of service of the roadway, AADT, and number of lanes. A higher value of congestion index indicates a higher level of congestion.

Variables selected for model development depended on the quality of the data provided, the purpose of the variables, and the significance of those variables in calculating the crash count. More than 6000 segments with an average length of 2.2 miles were identified over the six years of roadway data. MoDOT chose the beginning and ending points of the segments based on the geometric and traffic properties of the segments and were included in roadway segments database.

When a regressor is nearly a linear combination of other regressors in the model, the affected estimates are unstable and have high standard errors. This problem is called *collinearity* or *multicollinearity* (*31*). It is a good idea to find out which variables are nearly collinear with which other variables and remove them from the analysis. Two variables, "congestion index rate" and "pavement index," in the initial dataset were highly multicollinear with "congestion index" and "PSR" respectively. They were removed from the analysis. A multicollinearity diagnostic was conducted in SAS using PROC REG with the options COLLIN (*32*). Belsley et al. (*31*) suggest that in the results of collinearity diagnostics, when the value of 'condition index' is larger than 100, the estimates might have a fair amount of numerical error. The values of 'condition index' were found as 161.08 and 5210.034 for "pavement index" and ""congestion index rate" respectively.

5. RESULTS

Generalized linear model was used to model the crash counts on the Missouri interstate highway segments using a negative binomial link function. A summary of the parameter estimates and their standard errors for the NB models developed in this study are presented in Table 2. The results indicate that for almost all of the models the variables lane, width, shoulder width, and PSR were not statistically significant factors in crash occurrences.

The signs of the parameter estimates make sense: number of lanes has a negative sign for all models, indicating that higher number of travel lanes reduces the number of

www.manaraa.com

crashes. This is in contrast with some of the previous studies that found higher number of lanes associated with higher risk of crashes (*33-36*). They used both AADT/n, where n = number of lanes, and n in their studies. We used AADT and n. So, in our study, the coefficient of n stands for the effect of increasing the number of lanes while holding AADT constant for that segment. In other studies (e.g. Abdel-Aty and Radwan (*33*) and Milton and Mannering (*36*)), increasing n means increasing the total AADT for the segment. Therefore, the negative sign of the coefficient of n in our study implies that increasing the number of lanes while keeping AADT constant enhances safely. In other studies, increasing n implies that not only are we increasing the number of lanes, but we are also increasing the amount of traffic. Hence, the positive sign of the coefficient makes sense for these other studies. The natural logarithm of AADT has a positive sign for all models, which indicates a higher number of crashes with higher traffic volume.

Table 2. Parameter estimates and their standard errors for the different negative binomial models

Model Type	Intercept	No Lanes	Lane Width	Shoulder Width	LnAADT	Speed Limit	Congestion Index	PSR	Percent Commercial	Transition
Models for all collision types combined										
All severities (Φ=1.1777)	-8.6195 (0.7035)	-0.1483 (0.0276)	0.03 (0.0271)	*-0.0163* (0.0098)	1.2857 (0.0528)	-0.0416 (0.0036)	0.0275 (0.0407)	-0.0014 (0.0061)	-3.0067 (0.1981)	-0.1372 (0.0553)
Only fatal (Φ=1.7012)	-18.5471 (1.5354)	-0.2567 (0.0475)	-0.048 (0.0513)	0.0042 (0.0181)	1.7118 (0.1176)	-0.001 (0.0068)	0.1751 (0.0743)	0.0104 (0.0108)	-3.0752 (0.3890)	-0.4763 (0.0968)
Only nonfatal (Φ=1.1812)	-8.5948 (0.7052)	-0.1478 (0.0276)	0.0295 (0.0271)	*-0.0163* (0.0098)	1.2863 (0.0529)	-0.0419 (0.0036)	0.0259 (0.0407)	-0.0019 (0.0061)	-3.0193 (0.1984)	-0.1307 (0.0554)
Models for all severity levels combined										
Head on (Φ=9.5631)	-30.8784 (4.7084)	-0.5973 (0.1431)	0.0574 (0.1396)	0.0669 (0.0553)	2.6216 (0.3552)	-0.0048 (0.0212)	0.335 (0.2122)	-0.0217 (0.0315)	-1.1127 (1.1768)	-0.6067 (0.2968)
Rear end (Φ=1.7502)	-16.5226 (1.0727)	-0.3263 (0.0372)	-0.0088 (0.0348)	-0.0193 (0.0131)	2.0702 (0.0857)	-0.0449 (0.0048)	0.1544 (0.0556)	-0.0001 (0.0081)	-4.7338 (0.2825)	-0.3175 (0.0744)
Sideswipe same dir. (Φ=22.0498)	-30.6095 (7.8961)	-0.7841 (0.2326)	-0.2722 (0.3013)	-0.0652 (0.0816)	2.9805 (0.5859)	-0.0047 (0.0330)	0.239 (0.3256)	0.0161 (0.0527)	*-3.5892* (2.0097)	-0.5271 (0.4692)
Sideswipe opposite dir. (Φ=1.4712)	-23.9352 (1.2394)	-0.5014 (0.0395)	0.0085 (0.0348)	-0.0137 (0.0136)	2.669 (0.1007)	-0.0483 (0.0052)	0.4197 (0.0581)	-0.0007 (0.0080)	-3.4065 (0.3146)	-0.2656 (0.0765)
Angle (Φ=1.5897)	-23.2943 (1.4463)	-0.3038 (0.0447)	-0.0198 (0.0416)	-0.0241 (0.0157)	2.351 (0.1164)	-0.0276 (0.0061)	0.3661 (0.0683)	0.0095 (0.0093)	-3.1084 (0.3733)	-0.2812 (0.0875)

- Bold numbers indicate significance at 95% confidence level, and italic numbers at 90% confidence level.

- Φ represents the estimated dispersion parameter.

www.manaraa.com

Speed limit has a negative sign for all the models developed; indicating higher speed limits decrease the number of crashes. The sign can be explained as: these models do not indicate if the crash happened in an urban or rural area; it is therefore reasonable to state that fewer crashes occur in the rural areas as a result of lesser traffic, and rural areas have higher speed limit. The speed limit is another way to capture the changes in the number of crashes as a result of a change in type of area. The congestion index was also found to have a positive sign on models where it is a significant factor. This indicates that a higher number of crashes occur with more congestion, which is very similar to AADT. Percent commercial has a negative sign and was found to be significant, which indicates that higher percentage of heavy vehicles in the traffic mix results in fewer crashes. This indicates that drivers in general take caution around heavy vehicles. It was also found that the percentage of heavy vehicles had the highest effect on the reduction of rear-end crashes.

The transition variable was designated in the model to capture the effects of the safety strategies during the years 2005 through 2007. This factor was found to be statistically significant at 95% level of confidence and have a negative sign on all the models developed. The negative sign of the estimate indicates a reduction in the number of crashes during the implementation years, 2005 - 2007. The estimated values for this parameter indicates that the safety improvement strategies were mostly effective in reducing the fatal crashes compared to nonfatal crashes and in reducing the head-on crashes (leading cause of fatal crashes), compared to the other types of collisions (see the spider chart in Figure 1). The effect on crash type sideswipe-same direction is not shown in the figure as it was not found to be significant. A clear connection between the two findings can be observed from Figure 1; head-on collisions are the most severe types of crashes that result in fatalities.

The transition variable was used with four continuous quantitative levels from 0 before 2005, and then 0.25 to 0.75 from 2005 to 2007 for each year. It was used to investigate the predicted values of crashes in 2008, assuming with/without safety improvements, and the predicted numbers for different models were compared. shows the mean, standard deviation, min, max, and sum of the predicted crash counts in an interstate roadway segment for the year 2008, assuming there were/were not safety

www.manaraa.com

improvements implemented on the interstate highways. Comparing the "without" condition with the "with" condition, a drop can be observed in all the measures shown in Figure 2 presents a clear illustration of the percent reduction in the expected value of the number of crashes for 2008 as a result of the safety improvement program. It can be observed that the highest reduction (highest safety improvement effect) was 30% for only-fatal crashes. In terms of the collision type, the safety enhancement strategies had the highest effect on head-on crashes. This type of crash specifically results in high fatalities and the goal of the MSHSP was to reduce the number of fatal crashes. It was also found that the highway safety improvements result in a reduction of 18-33% in the number of other collision types including rear-end, sideswipe same- and opposite-direction, and angle crashes.

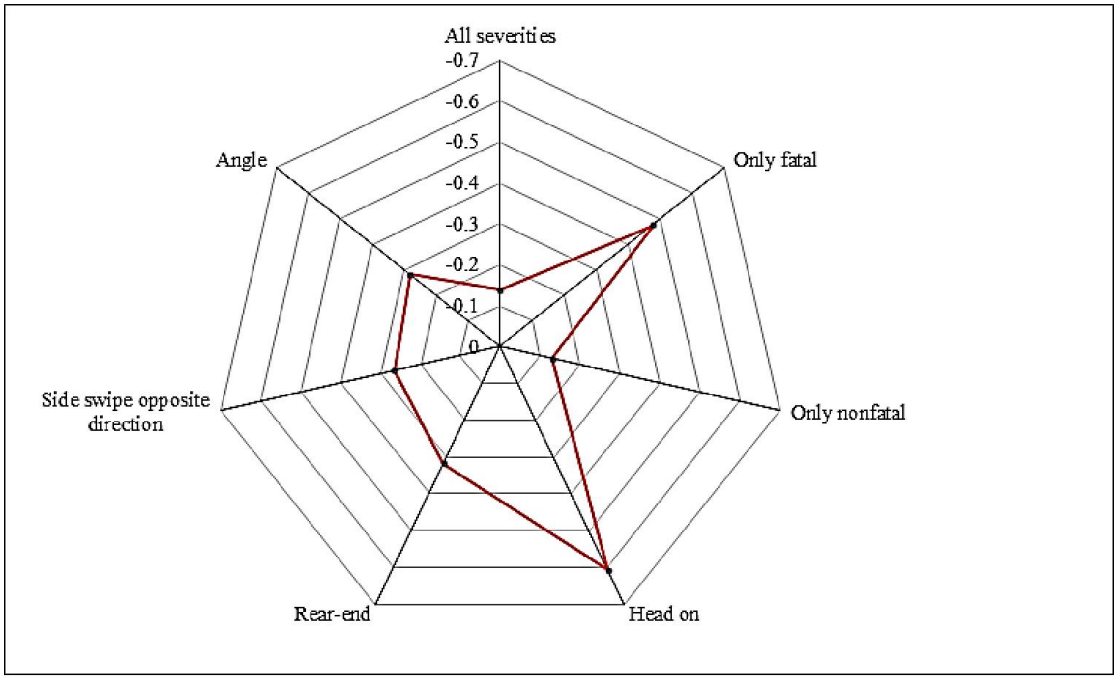

Figure 1. Graphical comparison of the effect of safety improvements on the crash models developed (values represent the estimate for the transition variable for each model).

Table 3, indicating the safety enhancing effects of the Missouri Blueprint strategies. It can also be observed that the maximum number of crashes included rear-end and sideswipe-same direction crashes. Figure 2 presents a clear illustration of the percent reduction in the expected value of the number of crashes for 2008 as a result of the safety improvement program. It can be observed that the highest reduction (highest safety

www.manaraa.com

improvement effect) was 30% for only-fatal crashes. In terms of the collision type, the safety enhancement strategies had the highest effect on head-on crashes. This type of crash specifically results in high fatalities and the goal of the MSHSP was to reduce the number of fatal crashes. It was also found that the highway safety improvements result in a reduction of 18-33% in the number of other collision types including rear-end, sideswipe same- and opposite-direction, and angle crashes.

Table 3. Comparison of the predicted crash count properties for 2008 with/without safety improvements

Model Type	Mean without*	Mean with*	Stdev without	Stdev with	Min without	Min with	Max without	Max with	Sum without	Sum with
Models for all collision types combined										
All severities	9.590	8.652	23.186	20.919	0.04815	0.04344	188.175	169.774	10453.12	9430.98
Only fatal	0.148	0.103	0.163	0.114	0.00083	0.00058	0.986	0.689	161.35	112.88
Only nonfatal	9.576	8.682	23.317	21.139	0.04706	0.04266	189.527	171.830	10438.02	9463.39
Models for all severity levels combined										
Head on	0.011	0.007	0.017	0.011	9.46E-06	6.00E-06	0.168	0.106	12.90	8.18
Rear end	5.668	4.467	16.520	13.020	0.001695	0.001335	183.230	144.405	6178.20	4869.07
Sideswipe same dir.	0.005	0.003	0.009	0.006	7.56E-07	5.09E-07	0.100	0.067	5.96	4.01
Sideswipe opp. dir.	2.325	1.905	7.629	6.251	0.000216	0.000177	114.669	93.958	2535.22	2077.33
Angle	0.445	0.361	0.949	0.768	0.00021	0.00017	10.942	8.861	486.10	393.67

* "without" and "with" indicates that model estimates for 2008 are determined assuming without and with safety improvements respectively

www.manaraa.com

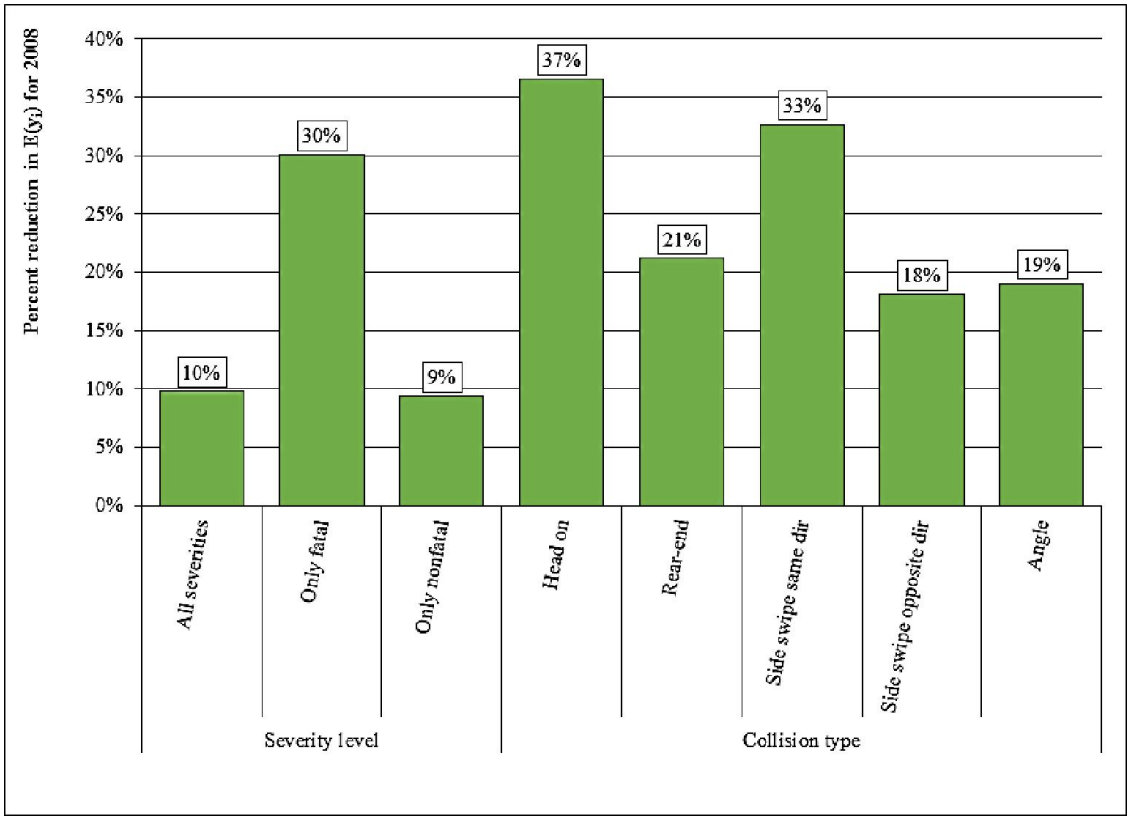

Figure 2. Percent reduction in the expected number of crashes predicted in 2008 as a result of safety improvement strategies.

6. CONCLUSIONS AND RECOMMENDATIONS

The objective of this study was to use a simple new approach to evaluate the effects of Missouri's Strategic Highway Safety Plan (MSHSP) on the number of crashes that occurred on the Missouri Interstate highways. Through the years 2004 to 2007, the MSHSP was implemented in enforcement, education, engineering, and public policy. Using a continuous variable through the implementation years, negative binomial regression models were developed and used to estimate the expected number of crashes in 2008 with and without the implementation of MSHSP. The results show that this safety enhancement program was able to reach its primary goal, i.e. to reduce the number and severity of serious injury crash types.

The study found a significant reduction of 10% for all crash severities combined and 30% for only fatal crashes. These strategies had the highest effect on the fatal crashes and particularly on the head-on crashes that result the most fatalities (*1*, *2*). It was also

www.manaraa.com

found that the highway safety improvements result in a reduction of 18-37% in the number of different collision types. The results from the model indicate that the MSHSP was a successful policy in reducing the overall number of crashes and decreasing the fatalities by decreasing the most severe injury crash types. The results are also consistent with many international studies and suggest that the safety strategic plans should be promoted as an effective treatment for highway crash fatalities (*37*, *38*). However, further analysis of particular SHSP implementation effectiveness that focus on the specific emphasis areas identified in the SHSP is warranted in future studies to obtain a more detailed understanding of how the implementation of specific safety measures affect safety. Provided the specific implementation data on the highways are available, future studies will consider examination of the effect of safety improvement plans (such as 'adding median barrier') on the type and injury severity of crashes.

7. REFERENCES

1. MoDOT. *Missouri's blueprint for safer roadways*. 2004 http://www.ite.org/safety/stateprograms/Missouri_SHSP.pdf. Accessed Nov. 10, 2013.
2. MoDOT. *Missouri's blueprint to arrive alive*. 2008 http://www.savemolives.com/documents/FINALBlueprintdocument.pdf. Accessed Nov. 10, 2013.
3. Geedipally, S.R., D. Lord, and S.S. Dhavala, *The negative binomial-Lindley generalized linear model: Characteristics and application using crash data.* Accident Analysis & Prevention, 2012. 45: p. 258-265.
4. Lord, D. and F. Mannering, *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives.* Transportation Research Part A: Policy and Practice, 2010. 44(5): p. 291-305.
5. Zou, Y., et al. *Comparison of Sichel and Negative Binomial Models in Estimating Empirical Bayes Estimates*. in *Transportation Research Board 92nd Annual Meeting*. 2013.
6. Shankar, V.N., F. Mannering, and W. Barfield, *Effect of roadway geometrics and environmental factors on rural freeway accident frequencies.* Accident Analysis & Prevention, 1995. 27(3): p. 371-389.
7. Hauer, E. and B. Persaud, *How to estimate the safety of rail-highway grade crossings and the safety effects of warning devices*1987: Transportation Research Board.
8. Knuiman, M.W., F.M. Council, and D.W. Reinfurt, *Association of median width and highway accident rates.* Transportation Research Record, 1993: p. 70-70.

www.manaraa.com

9. Squires, C.A. and P.S. Parsonson, *Accident comparison of raised median and two-way left-turn lane median treatments.* Transportation Research Record, 1989. 1239: p. 30-40.
10. Zegeer, C.V. and J.A. Deacon, *Effect of lane width, shoulder width, and shoulder type on highway safety.* State-of-the-Art Report, 1987(6).
11. Poch, M. and F. Mannering, *Negative binomial analysis of intersection-accident frequencies.* Journal of transportation engineering, 1996. 122(2): p. 105-113.
12. Datta, T.K., K. Schattler, and S. Datta, *Red light violations and crashes at urban intersections.* Transportation Research Record: Journal of the Transportation Research Board, 2000. 1734(1): p. 52-58.
13. Hauer, E., *Observational Before/After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*1997.
14. Persaud, B., R.A. Retting, and C.A. Lyon, *Crash reduction following installation of centerline rumble strips on rural two-lane roads.* Accident Analysis & Prevention, 2004. 36(6): p. 1073-1079.
15. Yu, R., M. Abdel-Aty, and M. Ahmed, *Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors.* Accident Analysis & Prevention, 2013. 50: p. 371-376.
16. Guo, F., X. Wang, and M.A. Abdel-Aty, *Modeling signalized intersection safety with corridor-level spatial correlations.* Accident Analysis & Prevention, 2010. 42(1): p. 84-92.
17. Miaou, S.-P., J.J. Song, and B.K. Mallick, *Roadway traffic crash mapping: A space-time modeling approach.* Journal of Transportation and Statistics, 2003. 6: p. 33-58.
18. Ahmed, M., et al., *Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway.* Accident Analysis & Prevention, 2011. 43(4): p. 1581-1589.
19. Shively, T.S., K. Kockelman, and P. Damien, *A Bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics.* Transportation research part B: methodological, 2010. 44(5): p. 699-715.
20. Huang, H. and M. Abdel-Aty, *Multilevel data and Bayesian analysis in traffic safety.* Accident Analysis & Prevention, 2010. 42(6): p. 1556-1565.
21. Persaud, B. and C. Lyon, *Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions.* Accident Analysis & Prevention, 2007. 39(3): p. 546-555.
22. Lord, D., S.P. Washington, and J.N. Ivan, *Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory.* Accident analysis and prevention, 2005. 37(1): p. 35-46.
23. Jones, B., L. Janssen, and F. Mannering, *Analysis of the frequency and duration of freeway accidents in Seattle.* Accident Analysis & Prevention, 1991. 23(4): p. 239-255.
24. Miaou, S.-P., *The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions.* Accident Analysis & Prevention, 1994. 26(4): p. 471-482.

www.manaraa.com

25. Shankar, V.N., J. Milton, and F. Mannering, *Modeling accident frequencies as zero-altered probability processes: an empirical inquiry.* Accident Analysis & Prevention, 1997. 29(6): p. 829-837.
26. Uhm, T., M.V. Chitturi, and A.R. Bill. *Comparing Statistical Methods for Analyzing Crash Frequencies*. in *Transportation Research Board 91st Annual Meeting*. 2012.
27. Kulmala, R., *Safety at rural three- and four-arm junctions: development and applications of accident prediction models.*, 1995, Technical Research Centre of Finland: Espoo. Finland.
28. Lee, J. and F. Mannering, *Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis.* Accident Analysis & Prevention, 2002. 34(2): p. 149-161.
29. ASTM-D6433-07, *Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys*, 2007, ASTM International: West Conshohocken, PA.
30. ASTM-E1489-08, *Standard Practice for Computing Ride Number of Roads from Longitudinal Profile Measurements Made by an Inertial Profile Measuring Device*, 2008, ASTM International: West Conshohocken, PA.
31. Belsley, D.A., E. Kuh, and R.E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. Vol. 571. 2005: John Wiley & Sons.
32. Littell, R.C., W.W. Stroup, and R.J. Freund, *SAS for linear models.* 2002: SAS Institute.
33. Abdel-Aty, M.A. and A.E. Radwan, *Modeling traffic accident occurrence and involvement.* Accident Analysis & Prevention, 2000. 32(5): p. 633-642.
34. Chang, L.-Y., *Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network.* Safety science, 2005. 43(8): p. 541-557.
35. Noland, R.B. and L. Oh, *The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data.* Accident Analysis & Prevention, 2004. 36(4): p. 525-532.
36. Milton, J.C. and F.L. Mannering, *The relationship between highway geometrics, traffic related elements, and motor vehicle accidents.* 1996.
37. Jung, S., Q. Xiao, and Y. Yoon, *Evaluation of motorcycle safety strategies using the severity of injuries.* Accident Analysis & Prevention, 2013. 59: p. 357-364.
38. Kempton, W., et al., *California Strategic Highway Safety Plan, Version 2. California Business, Transportation, Housing Agency Contributing Departments, Sacramento, CA.* , 2006.

www.manaraa.com

II. CRASH FREQUENCY MODELING USING NEGATIVE BINOMIAL MODELS: AN APPLICATION OF GENERALIZED ESTIMATING EQUATION TO LONGITUDINAL DATA

ABSTRACT

The prediction of crash frequency models can be improved when several years of crash data are utilized, instead of three to five years of data most commonly used in research. Crash data, however, generates multiple observations over the years that can be correlated. This temporal correlation affects the estimated coefficients and their variances in commonly used crash frequency models (such as negative binomial (NB), Poisson models, and their generalized forms). Despite the obvious temporal correlation of crashes, research analyses of such correlation have been limited and the consequences of this omission are not completely known. The objective of this study is to explore the effects of temporal correlation in crash frequency models at the highway segment level.

In this paper, a negative binomial model has been developed using a generalized estimating equation (GEE) procedure that incorporates the temporal correlations amongst yearly crash counts. The longitudinal model employs an autoregressive correlation structure to compare to the more traditional NB model, which uses a Maximum Likelihood Estimation (MLE) method that cannot accommodate temporal correlations. The GEE model with temporal correlation was found to be superior compared to the MLE model, as it does not underestimate the variance in the coefficient estimates, and it provides more accurate and less biased estimates. Furthermore, an autoregressive correlation structure was found to be an appropriate structure for longitudinal type of data used in this study. Ten years (2002 - 2011) of Missouri Interstate highway crash data have been utilized in this paper. The crash data comprises of traffic characteristics and geometric properties of highway segments.

Keywords: generalized estimation equation, longitudinal analysis, temporal correlation, crash frequency model, autocorrelation, autoregressive

www.manaraa.com

1. INTRODUCTION

Crash analysis research in general has focused on the estimation of traditional crash prediction models such as negative binomial (NB) and Poisson regression models and their generalized forms due to their relatively good fit to the crash (Shankar et al., 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Savolainen and Tarko, 2005; Mojtaba Ale Mohammadi et al., 2014a). Such crash prediction models take into account the crash frequency of a transportation facility (unit of analysis), such as an intersection or highway segment as a function of traffic flow and other crash-related factors. In these predictions, a greater amount of crash data, i.e. years of data, adds up to the reliability of the model estimates by reducing the standard errors (Lord and Persaud, 2000); However, the same unit generates multiple observations over the years that might be correlated due to unobserved effects related to specific entities that remain constant over time (Park and Lord, 2009; Castro et al., 2012; Bhat et al., 2014; Mannering and Bhat, 2014; Zou et al., 2014). In fact, these unobserved effects create a serial correlation in the repeated observations from the same unit over the years. Serial correlation in longitudinal data is an important issue, as it violates the independence assumptions on unobserved error terms in Poisson and/or NB crash frequency models, and creates inefficiency in the coefficient estimations and bias (underestimation) in estimation of standard error (Ulfarsson and Shankar, 2003; Washington et al., 2011; Dupont et al., 2013; Mohammadi et al., 2013; Bhat et al., 2014; Xiong et al., 2014).

Marginal models appear to be the most appropriate models for handling the temporal correlation, such as the work of Maher and Summersgill (1996) that uses an iterative solution based on the method of "constructed variables" presented by McCullagh and Nelder (1989). However, the extent and type of temporal correlation requires prior information that is not always known to the analyst (Lord and Persaud, 2000). Ulfarsson and Shankar (2003) tried to address the unit-specific serial correlation issue by using negative multinomial (NM) models in panel data and comparing the results with NB and random-effect negative binomial (RENB) model estimates. They showed that when there is correlation in the segment specific observations, the NM model is a much better fit compared to NB and RENB models. Dong, Richards, et al.

www.manaraa.com

(2014) developed multivariate random parameter models to account for the correlated crash frequency data as a result of unobserved heterogeneity. However, the model estimation methodology –the full Bayesian method– is complex, and the implementation and transferability of the method is not straightforward. Other research studies have been conducted in road safety analysis to account for such correlations in longitudinal data, yet consequences of the omission of the serial correlation are still not completely known. The most recent studies using longitudinal crash data include the work conducted by Venkataraman et al. (2014) to develop random parameter negative binomial models to investigate heterogeneity in crash means and the effects of interchange type on crash frequency.

Negative binomial models with a trend variable have also been used to study crash data with temporal correlation (Lord and Persaud, 2000; Noland et al., 2008; Quddus, 2008; Chi et al., 2012). Wang and Abdel-Aty (2006) used the technique of generalized estimating equations (GEE) to model rear-end crash frequencies at signalized intersections in order to account for the temporal and/or spatial correlation. GEE treats each highway segment as a cluster whose crash frequency observations have a temporal correlation over multiple years. In statistical terms, GEE captures the correlation incorporated in the error terms for model estimation. Hanley et al. (2003) showed that the use of GEE has the advantage of producing reasonably accurate standard errors and confidence intervals, especially when there are many subjects and few events. Hutchings et al. (2003) compared the performance of GEE with logistic regression by examining the change in parameter and variance estimates and the statistical significance of the independent variables. They found a lower number of significant variables when using the GEE method, and so recommended the use of nested structure models and GEE for analyzing motor vehicle crashes. H. L. Chang et al. (2006) applied the GEE procedure in a study of the effectiveness of drivers' license revocation and its impact on offenders in Taiwan. Lenguerrand et al. (2006) used multilevel logistic models (MLM), GEE, and logistic models (LM) to analyze hierarchical correlated crash data structure and found that both GEE and LM underestimate the parameters and confidence intervals, making MLM the most efficient model followed by GEE and LM models.

www.manaraa.com

Lord and Mahlawat (2009) used GEE method with an autoregressive (AR) correlation structure to investigate the effect of a small sample size and low mean value of crash frequency on the reliability of the inverse dispersion parameter estimate. They found that the standard errors of the models' coefficients are larger when the serial correlation is accounted for in the modeling process. Méndez et al. (2010) used both logistic regression and GEE models (with exchangeable correlation structure) to study the relationship of a car's registration year and its crashworthiness. Peng et al. (2012) also utilized the GEE method with an exchangeable correlation structure to study the relationship between drivers' inattention and their inability in lane keeping. Stavrinos et al. (2013) used a GEE Poisson regression to study the impact of various distractions on driving behavior. Since the GEE models are not based on maximum likelihood estimation (MLE), they used a Chi-square test to estimate the significance of the variables. Giuffrè et al. (2013) studied the concepts of dispersion and correlation in yearly crash frequency data and presented a quasi-Poisson model in a GEE framework to incorporate both the dispersion and temporal correlation. In comparing the GEE with the COM-Poisson regression model, they recommended the use of GEE whenever it is handy. GEE procedure is robust against misspecification of the correlation structure in the response variable, but in that case, one may lose significant model efficiency and cause a misleading interpretation of the results, which in turn affects the reliability of the final safety estimation (Giuffrè et al., 2013).

The examples outlined above illustrate how GEE is not actually a regression model, but rather a method used to estimate models for data characterized by serial correlation. Throughout this paper, the models with temporal correlation that use GEE procedure are referred to as the GEE models. Unlike the traditional marginal models, the GEE models can handle temporal or other forms of correlation, even if the extent and type of correlation is unknown. However, Giuffrè et al. (2007) demonstrated that utilizing data correlation structure in safety modeling results in higher estimation precision. Although they have acknowledged that GEE models generally are robust to misspecification of the correlation structure (Liang and Zeger, 1986), and researchers believe the true correlation structure is important only when marginal models are estimated by using data with missing values, but when the specified structure does not

www.manaraa.com

incorporate all of the information on the correlation of measurements within the subjects, loss of efficiency in estimates can be expected Ballinger (2004).

The objective of this paper is to present an application of GEE for developing a longitudinal negative binomial model that incorporates the temporal correlation of repeatedly measured crash counts over 10 years (2002-2011). For this purpose, the presence of an AR correlation structure, (AR(1), i.e. autoregressive with lag 1) in the longitudinal data was first determined by the Durbin-Watson test and then validated by the results. In this paper, a traditional NB model, namely a *MLE model,* was also developed and the results were compared with the GEE estimation results. This approach, however, assumes that there is no unaccounted unobserved heterogeneity correlated with crash-related covariates that creates a fake autoregressive correlation among the observed crash frequencies over the years. The remainder of the paper presents the technique of GEE approach followed by description of the crash data used. Results and findings are then followed by conclusions of the study.

2. METHODOLOGY

To measure the influence of different factors that change every year, crash data was grouped into clusters (each highway segment acts as a cluster), with crash frequency observations made over time in the same cluster tending to be more alike than observations across clusters. That means a segment is a cluster within which the crash frequencies are correlated over several years. This temporal correlation creates difficulties for traditional frequency model estimations (Ulfarsson and Shankar, 2003; Mannering and Bhat, 2014). While standard maximum-likelihood analysis specifies the full conditional distribution of the dependent variable, quasi-likelihood analysis postulates a relationship between the expected value of the response variable (crash frequency), the covariates, and between the conditional mean and variance of the response variable (Gill, 2001; Zorn, 2001). GEE is classified as a multinomial analogue of a quasi-likelihood function that offers different approaches to handle serial correlations (see Fitzmaurice et al. (1993) for details).

www.manaraa.com

Zeger and Liang (1986) first used the GEE technique by extending the approach of generalized linear model to correlated data in the context of repeated observations over time. Consider a model of crash frequency observations at a highway segment i during time t (Y_{it}) and k covariates (X_{it}), where i indexes the N clusters (highway segments) and t indexes the T repeated measurements (time points), a function h can be defined to specify the relationship between Y_i and X_i (Zorn, 2001):

$$\mu_i = E(Y_i) = h(X_i\beta) \tag{1}$$

where,

μ_i: expected value of the crash frequency at segment i, (Y_i), $i = 1, 2, \ldots, N$

β: $k \times 1$ vector of estimable parameters

X_i: $t \times k$ matrix of covariates for segment i $(i = 1, 2, \ldots N, t = 1, 2, \ldots T)$.

The variance of Y_i is specified as a function g of the mean μ_i:

$$V_i = g(\mu_i)/\phi \tag{2}$$

where,

V_i: variance of Y_i and

ϕ: scale parameter.

The quasi-likelihood estimate of β is then the solution to a set of k "quasi-score" differential equations (Zeger and Liang, 1986; Zorn, 2001):

$$U_k(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1}(Y_i - \mu_i) = 0 \tag{3}$$

where,

$D_i = \mu_i/\beta, V_i = \frac{(A_i)^{1/2} R_i(\alpha)(A_i)^{1/2}}{\phi}$

A_i: $T \times T$ diagonal matrices with $g(\mu_{it})$ as the t^{th} diagonal element,

$R_i(\alpha)$: a $T \times T$ matrix of the working correlations across time for a given Y_i, and

α: a vector of unknown parameters with a specific structure (according to the type of correlation structure).

The GEE estimator results can be obtained by substituting Equation (4) into Equation (3). In the resulting equation, it can be seen that GEE is an extension of the generalized linear model (GLM) approach, and that it reduces to the GLM when T equals

www.manaraa.com

1 (Zorn, 2001). To solve GEE, every element of the correlation matrix R_i is required to be known; although, it is not always possible to know the exact correlation type for the repeated measurements. To overcome this issue, the use of a "working" matrix $\hat{V}$ for the correlation matrix V_i based on the correlation matrix $\widehat{R}_i$ was proposed by Liang and Zeger (1986). The estimate of β is then found from the following differential equation:

$$U_k(\beta) = \sum_{i=1}^{N} D_i' \hat{V}_i^{-1} (Y_i - \mu_i) = 0 \tag{4}$$

The covariance matrix of β is given by

$$cov(\hat{\beta}) = \sigma^2 \left[\sum_{i=1}^{N} D_i' \hat{V}_i^{-1} D_i\right]^{-1} \left[\sum_{i=1}^{N} D_i' \hat{V}_i^{-1} V_i \hat{V}_i^{-1} D_i\right] \left[\sum_{i=1}^{N} D_i' \hat{V}_i^{-1} D_i\right]^{-1} \tag{5}$$

Equations 5 and 6 provide almost always consistent estimates of β even with an inadequate estimate of the correlation matrix V_i. Therefore, the confidence interval for β will be correct and there is no need to know the type of temporal correlation, even when the covariance matrix is specified incorrectly. However, to assume that $\hat{\beta}$ is an accurate estimate of β, the observation for each roadway segment should be known with no missing observations, otherwise, it will result in biased coefficient estimates (Lord and Persaud, 2000).

The potential positive autocorrelation in crash frequency data was examined by the Durbin-Watson (DW) test. This test statistically examines if the residuals from a regression model are independent. The null hypothesis is that there is no autocorrelation ($\rho = 0$), and the alternate hypothesis is that the autocorrelation is positive ($\rho > 0$). The test statistic can be calculated as:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{6}$$

Where, $e_i = y_i - \hat{y}_i$ and y_i and $\hat{y}_i$ are the observed and predicted values of the response variable for segment i, respectively. The d statistic becomes smaller as the serial correlations increase. For the crash data set used, the DW test statistic was found to be 0.5235 with 6849 degrees of freedom for the order 1 lag of autocorrelation and the null hypothesis was rejected at a significance level of 0.01, indicating the presence of a

www.manaraa.com

positive autocorrelation. As mentioned earlier, what is perceived as temporal correlation in the crash data might be induced (infected) by the effects of unobserved heterogeneities that vary across years with a slow trend (Ulfarsson and Shankar, 2003). In this study, all the covariates in the models are considered at the highway segment level and assumed that there are no unobserved heterogeneity effects on the considered covariates of the model. This is a true assumption if the values of the working correlation matrix across time do not change drastically with a change in the analysis time period. In this study, different analysis periods were considered to determine the correlation matrices: A) three years (2002 – 2004), B) seven years (2002 – 2008), and C) ten years (2002 – 2011). Obtaining similar values of serial correlation from these analysis periods will provide evidence of the existence of a disinfected temporal correlation that can be addressed by the GEE method. In such conditions, one expects that the values of the parameter estimates and their level of significance will also be very similar to each other, no matter the number of years of data used for the longitudinal analysis. Table 1 shows the working correlation matrices for the AR correlation structure and the value of working correlation assuming an exchangeable correlation structure. The 1st to 3rd rows in each column of the table shows the AR correlations (lag 0 to lag 3) for the three time periods, respectively. It can be observed that the correlation values for each lag are very similar to each other. This indicates that an AR correlation structure can be reliably used to address the temporal serial correlation.

Table 1. Correlation values for the autoregressive Type 1 and exchangeable structure

A) 3 years, 2002-2004, Exchangeable working correlation = 0.7245

B) 7 years, 2002-2008, Exchangeable working correlation = 0.7651

C) 10 years, 2002-2011, Exchangeable working correlation = 0.7552

Working Correlation Matrix for autoregressive type 1 correlation structure				
	Lag0	**Lag1**	**Lag2**	**Lag3**
Period A	1	0.7462	0.5569	N/A
Period B	1	0.7836	0.6140	0.4811
Period C	1	0.7537	0.5681	0.4281

N/A indicates non-applicability of the AR correlation for the lag in corresponding column for that analysis period

www.manaraa.com

In this paper, an AR(1) (autoregressive with lag 1) correlation structure was used in the GEE procedure (Zorn, 2001; Allison, 2012). This correlation structure indicates that two observations within a segment made close in time tend to be more correlated than two observations made far apart in time from the same segment. There are other correlation structures, such as an exchangeable structure that specifies, for each segment, the temporal correlations are equal across the years. And an independent structure that forces the cross-time correlation to be zero for each segment for which, the GEE estimation reduces to ordinary MLE, so the estimated coefficients would be the same as those for traditional NB model. For detailed information about various correlation structures, refer to the work published by Hardin and Hilbe (2007).

The GENMOD procedure with a REPEATED option in SAS V.9.3 was used to follow the GEE procedure and develop the model of interest (SAS, 2008; Allison, 2012). Two goodness-of-fit indices --quasi-likelihood under the independent model criterion (QIC), and its sample version, called $QICu$-- were also found to determine the reliability of the coefficients estimates. As GEE is a quasi-likelihood-based method, Pan (2001) suggested using the QIC which is equivalent to the AIC in evaluating competitive models' fit. QIC is defined as

$$QIC(R) = -2Q\left(\hat{\beta}(R), \phi\right) + 2trace(\hat{\Omega}_I \hat{V}_R) \tag{7}$$

where, $Q\left(\hat{\beta}(R), \phi\right)$ is the quasi-likelihood function under the independent working correlation assumption, evaluated with the parameter estimates under the working correlation of interest R, $\hat{\beta}(R)$, $\hat{\Omega}_I$ is the inverse of the model-based covariance estimate and $\hat{V}_R$ is the robust covariance estimate. The underlying principle of QIC is comparable to AIC. The first term of QIC (refer to Equation 5) is the quasi-likelihood computed using a specified working correlation structure, which corresponds to the likelihood estimation equivalent of the AIC and likewise the second term is the penalty which serves similar effect as the second term in computing AIC (Hardin and Hilbe, 2007). Hardin and Hilbe (2007) also suggested the use of $QICu$ to approximate QIC. However, $QICu$ cannot be applied to select the working correlation matrix R, as the presumption of $QICu$ is that the specification of working correlation is correct. $QICu$ is defined as

www.manaraa.com

$$QIC_u(R) = -2Q\left(\hat{\beta}(R), \phi\right) + 2p \quad (8)$$

where, p is the number of regression parameters. Similar to the concept of AIC, the smaller the QIC and QICu are, the better the fit of the model. The importance of the above measures of goodness-of-fit is significant when comparing different models (e.g. with different correlation structures). This study also utilizes the chi-square values of the estimated model parameters and the cumulative residual (CURE) plots to investigate the quality of fit and compare the GEE models with the common negative binomial model (Hauer and Bamfo, 1997; Lord and Persaud, 2000).

3. CRASH DATA

3.1. Data Description

The Missouri Department of Transportation (DOT) portal of safety investigation provided access to the accident data base for all of the recorded years of data. The data consists of all levels of crash severity for motor-vehicle crashes (fatal to property-damage-only accidents) at 17 interstate highways in the state of Missouri from 2002 to 2011. Table 2 presents the major yearly characteristics of these highways. The highways with a total length of about 1200 miles were classified as divided highways (65% in rural areas and 35% in urban areas). The total number of crashes in the data set analyzed was 167,783 crashes, out of which 37% occurred in rural areas and 63% in urban areas. The rate of crash (per mile, per vehicle) for a segment in each year is shown in Column 2 of Table 2, with the total number of crashes on all interstate highways presented in Column 3.

The initial explanatory variables considered for this analysis were the area type (urban or rural), number of lanes, lane width (min of 10 ft. to max of 18 ft.), shoulder width (min 0 ft. to max of 15 ft.), AADT (Annual Average Daily Traffic), speed limit (55, 60, 65, and 70 mph), PSR (pavement serviceability rate), PCI (pavement condition index), CIR (congestion index rate), and percentage of commercial vehicles (truck percentage). These variables were selected for model development depending on the quality of the data provided, function of the variables, and the significance of those variables in calculating the crash frequency.

www.manaraa.com

PSR is equal to two times the Ride number plus the PCI. The Ride number is an index derived from controlled measurements of longitudinal profile in the wheel tracks correlated with rideability of a pavement using a scale of 0 to 5, with 5 being perfect and 0 being impassable. The PCI is a numerical rating of the pavement condition that ranges from 0 to 100, with 0 being the worst possible condition and 100 being the best possible condition. More information on the indices of the Ride number and PCI can be found on the standards ASTM D6433-07 (ASTM-D6433-07, 2007) and ASTM E1489-08 (ASTM-E1489-08, 2008), respectively. A higher value of PSR indicates a higher serviceability of the pavement. The CIR presents the congestion level, calculated by incorporating the level of service of the highway, AADT, and number of lanes. A higher value of this variable is a sign of a higher level of congestion.

Table 2. Descriptive statistics of segment properties of Missouri interstates (2002-2011)

Year	Number of crashes, per segment (min-mean-max)	Number of crashes, Total	Annual Average Daily Traffic (AADT) (min-mean-max)	Number of lanes (min-mean- max)	Pavement Serviceability Rate (min-mean-max)	Percent commercial (min-mean-max)
2002	0-17.5-347	18955	1985-29477-101594	2-2.6-7	19.3-32.1-66.4	0.041-0.215-0.582*
2003	0-18.1-361	19581	1867-29467-98485	2-2.6-7	17.4-32.3-37.4	0.041-0.208-0.406
2004	0-17.9-131	19343	1919-29861-109420	2-2.6-6	18.9-32.1-37.3	0.046-0.229-0.582
2005	0-17.5-150	19101	1865-29933-109573	2-2.6-6	24.0-33.0-39.6	0.045-0.234-0.582
2006	0-17.3-176	18922	1874-30418-114753	2-2.5-6	23.4-34.1-37.5	0.049-0.234-0.582
2007	0-19.0-168	19308	1893-31446-115901	2-2.6-6	22.9-34.1-37.6	0.049-0.229-0.622
2008	0-16.9-121	18474	1920-30301-115182	2-2.6-6	24.9-33.5-37.0	0.049-0.228-0.582
2009	0-17.3-133	17823	1955-30678-107689	2-2.7-6	26.3-33.4-37.0	0.034-0.234-0.582
2010	0-17.0-149	17900	830-30335-106612	2-2.7-6	18.4-30.7-36.8	0.050-0.230-0.674
2011	0-16.5-188	17742	813-30158-105546	2-2.7-6	19.9-31.1-37.5	0.050-0.224-0.674

* This high value of truck percentage probably represents night time at specific segments of the highways with low traffic

More than 6000 segments, with an average length of 2.2 miles, were identified over the 10 years of crash data. The Missouri DOT determined the segmentation, i.e., chose the beginning and ending points of the segments based on the homogeneity of the geometric (number of lanes, lane width, etc.) and traffic properties (AADT) of segments. Other segment properties that are recorded in the segmentation database take the value of that property that prevails throughout the majority of the segment. That is, for example, if the majority of a segment has pavement type A and the rest is type B, the value for the pavement type of that segment is recorded as type A. Therefore, one cannot say for certain that a segment is homogenous in terms of all the variables throughout the length

www.manaraa.com

of the segment. Some of the pavement-related variables that were in the original dataset include: shoulder type, surface type, PSR, and pavement index out of which, only the continuous variable PSR was considered in this study. If the geometric or traffic characteristic of a segment changed for any year, that segment was identified as a new segment with new sets of properties for the rest of the years until it undergoes another change in the deterministic homogenous properties. The crash related databases can be accessed through Missouri DOT's virtual private network that requires coordination with the transportation planning section.

Figure 1 depicts the total number of crashes occurred on the interstate highways of Missouri during 2002 to 2011. Only those highways with the most variation in crashes are shown in the chart. It can be observed that there is not much variation in the number of crash statistics over the years, which might imply there is correlation in crash frequency observations. Interstate-70, I-44, and I-270 have the highest total number of crashes per year amongst the highways. For similar studies that have used several years of data with consideration of the correlation amongst the repeated observations, the interested reader is referred to Guo et al. (2010a), Venkataraman et al. (2011), Castro et al. (2012), Venkataraman et al. (2013), and Venkataraman et al. (2014).

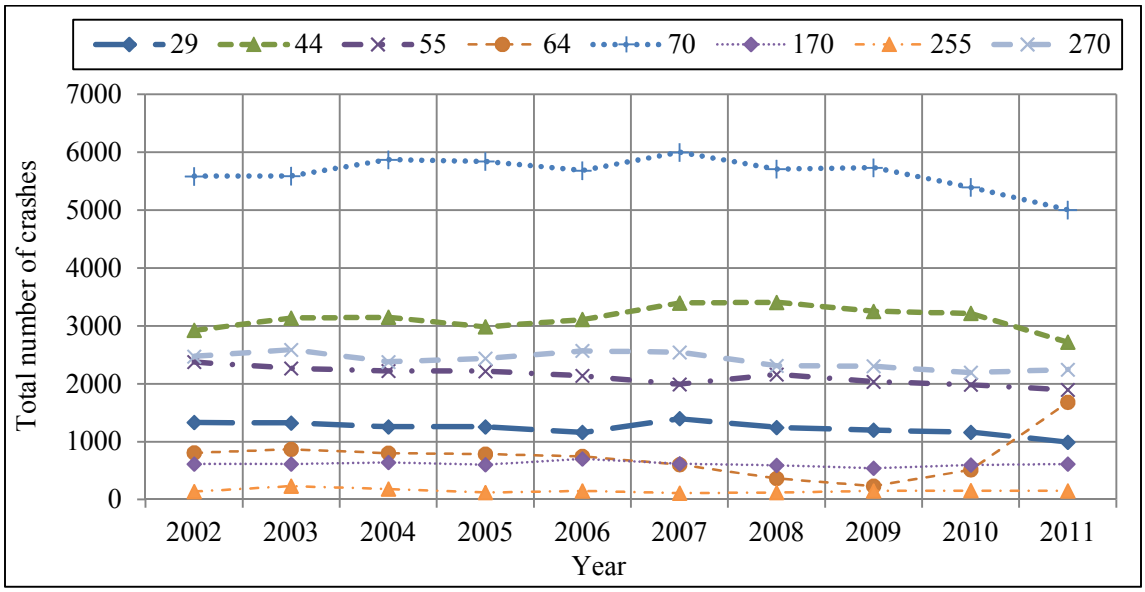

Figure 1. Total number of crashes on a selected few of the interstate highways of Missouri with most variation. (legend presents the name of the interstate highways, e.g. 44 indicates interstate 44)

www.manaraa.com

3.2. Multicollinearity

Variables that were involved in multicollinearity were removed from the analysis. When a regressor variable is nearly a linear combination of other regressors in the model, the affected estimates are unstable and have high standard errors. This problem is called *collinearity* or *multicollinearity*. It is beneficial to find out which sets of variables are multicollinear and withdraw one variable from each set (Washington et al., 2011). In this study, in addition to the Pearson's correlation coefficient, variance inflation factor (VIF), tolerance, and condition index (CI) were also used for detecting multicollinearity (Littell et al., 2002). Table 3 presents the Pearson's correlation coefficient between the suspected variables, VIF, and Tolerance values. The approach used in the analysis follows that of Belsley et al. (2005).

Table 3. Pearson correlation coefficients and collinearity diagnostics

Pearson Correlation Coefficients					**Collinearity Diagnostics**	
Parameter	PSR†	PCI†	CIR†	Percent commercial	VIF‡	Tolerance
PSR	1	**0.99994**	0.01968	**0.0976**	7.653	0.13066
PCI	**0.99994**	1	0.01959	**0.09764**	6.261	0.15971
CIR	0.01968	0.01959	1	**0.64054**	8745.4	0.0001143
Percent commercial	**0.0976**	**0.09764**	**0.64054**	1	8745.4	0.0001143

† PSR, PCI, and CIR stand for pavement serviceability rate, pavement condition index, and congestion index rate, respectively

‡ VIF represents the variance inflation factor

Note: Bold values are statistically significant at 99% level of confidence

The Pearson correlation coefficient between the PCI and PSR is almost 1. Also, CI values of 5210 and 161 were found in the analysis, indicating that there is a group of multicollinear variables in the dataset including the variable PCI. Belsley et al. (2005) suggest that, when this number is larger than 100, the estimates might have a fair amount of numerical error. The VIF value for CIR is as high as 8745, which is much higher than 10 and, according to Kutner et al. (2004), an indication of multicollinearity. Percent commercial also has a high VIF value which probably indicates collinearity with the CIR. The variables PCI and CIR were removed from the analysis and a multicollinearity diagnosis was rerun, with no sign of multicollinearity was observed.

www.manaraa.com

3.3. Sample Size

As this study uses the GEE method, the number of observations for each class of crash incorporating categorical factors within each year should be examined (Maas and Hox, 1999; Hutchings et al., 2003). For the sake of brevity, the frequency tables for each factor at each step of this process are not included in the paper. However, the detailed results of this examination for the variables shoulder width, PSR, number of lanes, and speed limit can be found elsewhere (Mohammadi, 2014).

In order to verify the sufficiency of the sample size for analysis, the variable shoulder width was examined within each year for the number of observations in each class. Assuming 60 observations as a sufficient number in each class (Mancl and DeRouen, 2001), it was observed that the shoulder width classes except 10 ft. (with 86% of the observations) lacked enough observations. The observations for the various classes of shoulder width were categorized into three groups according to its mean (9.4814) and standard deviation, SD (1.5622): A) less than mean – SD, B) between mean – SD and mean + SD, and C) more than mean + SD. This categorization also did not work as groups A and C still lacked enough data within the majority of the years considered. Further, the 25^{th}, 50^{th}, and 75^{th} percentiles (used to group observations in sufficient numbers) did not resolve the problem, as all three percentile values had the same value, 10 ft. Therefore, it was decided to remove the variable shoulder width from the analysis.

PSR, which is a factor related to the PCI, is a continuous variable and the only way to verify the sufficiency of observations was to categorize the observations according to its mean value (32.474) and standard deviation (2.640). In order to obtain an optimal categorization (with at least 60 observations within each year), three classes of PSR were defined in the following format:

- Class 1, "PSR < (mean – 0.3 x SD) = 24.55",
- Class 2, "(mean – 0.3 x SD) < PSR < (mean + 0.3 x SD)", and
- Class 3, "(mean + 0.3 x SD) < PSR".

The number of observations categorized by number of lanes was then tabulated to verify the sufficiency of observations in each class. It was observed that the categories equal to 4, 5, 6, and 7 lanes lacked enough observations. For this reason, it was decided to

www.manaraa.com

combine these groups into one category of number of lanes which include more than three (nolanes>3). Therefore, three groups of nolanes were determined as follows: nolanes=2, nolanes=3, and nolanes>3. Similar to the other variables, speed limit (SL) was also examined for the number of observations in each class per year and the inspection showed that few groupings (6 out of 40 groups with observations in the range of 41 to 53) lack enough observations for some years. The variable speed limit was nevertheless used in the model, as the shortage in the number of observations was relatively minor.

3.4. Confounding Effects and Variable Specification

The possible confounding effect of the classes of number of lanes, and speed limit (SL) with the variable area type (urban or rural) was examined. Table 4 presents the distribution of the number of observations for number of lanes and speed limit classes within each area type. The top and bottom row for each combination presents the actual number and percentage of the observations (e.g. 915 segments were observed in the urban areas with speed limit of 55 mph that consists 13.9% of all the observed segments in the dataset).

Table 4. Number and percentage of observations within area types, by number of lanes and speed limit

Area type	Number of Lanes				Speed Limit (mph)				
	2	3	>3	Total	55	60	65	70	Total
Rural	2463	39	-	2502	0	28	14	2459	2501
	36.2	0.57	-	36.8	0	0.43	0.21	37.35	37.99
Urban	1641	1474	1188	4303	915	1194	913	1060	4082
	24.1	21.7	17.5	63.2	13.9	18.14	13.87	16.1	62.01
Total	4104	1513	1188	6805	915	1222	927	3519	6583
	60.3	22.2	17.5	100	13.9	18.56	14.08	53.46	100

For rural areas, highways with only two or three lanes were observed in the data. Moreover, only 39 observations were found for number of lanes equal to three. Two analyses were conducted: 1) rural segments with three lanes were deleted from the dataset, and 2) the 39 rural segments with three lanes were combined with rural segments with two lanes rather than delete these observations. Since no significant change was

www.manaraa.com

observed in the estimates, it was decided to keep the 39 observations combined with the observations with two lanes. For urban areas, sufficient data were found for each class.

For the speed limit classes, it was observed that the rural area lacks a sufficient number of observations in the SL classes of 55, 60, and 65 mph, while 98.32% of the rural segments have a speed limit of 70 mph. This issue was not observed in the urban category. One might suspect that there is confounding amongst the variables speed limit, number of lanes, and area type. To investigate this issue, several different analyses were conducted using these newly defined categories to examine the effect of each variable when fitted simultaneously as classification variables. Since there were confounding effects, some of the effects and their interactions could not be estimated. Such estimability issues arising out of confounding are to be expected. As a solution, it was decided to define new dummy variables, and each represents one of the $area \times nolanes \times SL$ interactions. Three groupings were chosen for the three nolanes categories of two, three, and more than three lanes, four SL categories of 55, 60, 65, and 70 mph, with the exception that there was no observation for the rural areas with speed limit 55 mph across all categories of nolanes. Table 5 presents these dummy variables along with the number of observations for each category. The three parts of the dummy variables indicate the area type, number of lanes, and speed limit criteria, respectively. There are overall 12 categories defined for urban and six categories for the rural area segments.

The group in rural area with three lanes and speed limit of 60 mph (rural_3_60) had zero observations, and therefore was not used in the model. The rural category with two lanes and speed limit of 70 mph (rural_2_70) was used as the base condition in the model. The other categories in the rural area type did not have the target value of 60 observations, but were retained to avoid removing the data. The soundness of this decision was double-checked by running two models –one with and another without the small-sized rural variables– and comparison of the two models. All the dummy variables in the rural category were not found to be significant variables in the model. In other words, these categories did not result in statistically different effects from the base condition represented by rural_2_70. This might be because of the small number of

www.manaraa.com

observations that exist in those categories. Therefore, all of those rural categories were lumped together with the rural_2_70 group.

Table 5. List of the dummy variables considered for the analysis

Variable[1] (Obs.)[2]		Definition
Urban_2_55	(699)	1, if area = Urban, number of lanes = 2, and speed limit = 55 mph, 0 otherwise
Urban_2_60	(180)	1, if area = Urban, number of lanes = 2, and speed limit = 60 mph, 0 otherwise
Urban_2_65	(261)	1, if area = Urban, number of lanes = 2, and speed limit = 65 mph, 0 otherwise
Urban_2_70	(877)	1, if area = Urban, number of lanes = 2, and speed limit = 70 mph, 0 otherwise
Urban_3_55	(415)	1, if area = Urban, number of lanes = 3, and speed limit = 55 mph, 0 otherwise
Urban_3_60	(323)	1, if area = Urban, number of lanes = 3, and speed limit = 60 mph, 0 otherwise
Urban_3_65	(484)	1, if area = Urban, number of lanes = 3, and speed limit = 65 mph, 0 otherwise
Urban_3_70	(140)	1, if area = Urban, number of lanes = 3, and speed limit = 70 mph, 0 otherwise
Urban_3p_55[3]	(216)	1, if area = Urban, number of lanes > 3, and speed limit = 55 mph, 0 otherwise
Urban_3p_60	(691)	1, if area = Urban, number of lanes > 3, and speed limit = 60 mph, 0 otherwise
Urban_3p_65	(168)	1, if area = Urban, number of lanes > 3, and speed limit = 65 mph, 0 otherwise
Urban_3p_70	(143)	1, if area = Urban, number of lanes > 3, and speed limit = 70 mph, 0 otherwise
Rural_2_60	(28)	1, if area = Rural, number of lanes = 2, and speed limit = 60 mph, 0 otherwise
Rural_2_65	(12)	1, if area = Rural, number of lanes = 2, and speed limit = 65 mph, 0 otherwise
Rural_2_70[1]	(2422)	1, if area = Rural, number of lanes = 2, and speed limit = 70 mph, 0 otherwise
Rural_3_60	(0)	1, if area = Rural, number of lanes = 3, and speed limit = 60 mph, 0 otherwise
Rural_3_65	(2)	1, if area = Rural, number of lanes = 3, and speed limit = 65 mph, 0 otherwise
Rural_3_70	(37)	1, if area = Rural, number of lanes = 3, and speed limit = 70 mph, 0 otherwise

1. Dummy variables were defined in this format due to confounding effects of the incorporating variables. Variables in bold were used in the final model.

2. The values in the parentheses present the number of observations for the corresponding variable.

3. 3p means 3-plus indicating more than 3 lanes.

Finally, to incorporate the impact of the main variables on crash count differently in urban and rural road segments, a new dummy variable, "area", was defined and was set to be zero for rural (base category) and one for urban. The interactions of this variable with the other main factors of the model were considered to be included in the model as:

$$AREADT = Area \times lnAADT;$$
$$AREACOMMERCIAL = Area \times Percentcommercial;$$
$$AREAWIDTH = Area \times Lanewidth;$$

www.manaraa.com

$AREAPSR = Area \times PSR$;

The interaction term AREAWIDTH term gave rise to a complicated convergence iteration process that did not satisfy the convergence criterion. Investigating the number of observations per lane width category for the rural and urban areas revealed that, since only three of seven lane width classes were observed in the rural area, estimation of the complete set of interaction effects was not possible with the available data. Therefore, this term was removed from the model. The interaction term AREAPSR was not found to be statistically significant in any of the two GEE and MLE models and so was also removed from the analysis. However, the main factors that were not found to be statistically significant but were involved in a significant interaction term were left to remain in the model. That is, the "*conditional effects"* of the main factors that were not statistically significant remained in the model to correctly interpret the interaction parameters (Nelder, 1977; Cox, 1984). Although, the dummy variable "area" was not used in the model as it was confounded by the combinatory dummy variables (see Table 5). Table 6 presents the name and definition of the continuous and categorical variables used in this study.

Table 6. List of the continuous and classification variables considered for the analysis

Variable	**Definition**
Continuous variables	
LnAADT	Natural logarithm of the annual average daily traffic in vehicle per day.
Percentcommercial	The annual average percentage of trucks or heavy vehicles.
Lanewidth	The width of the highway lane in feet.
Areadt	The interaction between two variables "Area"† and "LnAADT".
Areacommercial	The interaction between two variables "Area" † and "Percentcommercial".
Classification variable	
PSRclass	Classification of PSR, an index for pavement serviceability rate: • PSRclass = 1 when PSR < 24.55 • PSRclass = 2 when 24.55 ≤ PSR < 40.39 • PSRclass = 3 when 40.39 ≤ PSR

† "Area"=0 for rural, and =1 for urban areas; though, it was not used in the model due to confounding effect with combinatory dummy variables (see Table 5)

www.manaraa.com

4. RESULTS AND DISCUSSION

4.1. Model estimates and comparisons

Two models were developed, an autoregressive type 1 (AR1) GEE model incorporating temporal correlation and a traditional NB (MLE model). A negative binomial distribution was used to specify the error structure; however, In the GEE model, the crash frequency for each year was used as a separate observation (a dependent variable), to be modeled by the crash covariates. Table 7 presents the results of the estimates and standard errors of the coefficients for the GEE and MLE models along with the QIC and $QICu$ values for the GEE model. The signs of the parameter estimates make sense, however, interpreting these signs may not be completely accurate, as some variables were not found to be significant in the GEE model. A primary objective of this study is also to find out whether factors that are found to be statistically significant in MLE model are truly significant.

The natural logarithm of AADT has a positive sign for both models that indicates a higher number of crashes with higher traffic volume. PSRclass was not found to be a statistically significant factor in any of the models, except lower classes of PSR in the MLE model. Observing the trend of the PSRclass estimates reveals that higher classes of PSR (better pavement) results in comparatively lower crash frequency. Truck percentage (percentcommercial) was not found to be a significant factor in the GEE model, but it was found to be significant with a negative value of estimate in the MLE model, indicating that more heavy vehicles result in fewer crashes. This indicates that drivers are generally more cautious when they see or are traveling close to large vehicles. According to Carson and Mannering (2001), the reduction in crash frequency due to this factor might also relate to the reduction in speed that heavy vehicles have on the traffic stream. Lao et al. (2014) also found similar results of the effect of truck percentage on rear-end crash occurrences.

Lanewidth was only found to be significant in the MLE model, with a positive sign indicating that a higher lane width increases the likelihood of crash occurrence, which may seem counterintuitive. (For example, see the works of Li et al. (2008), Manuel et al. (2014), who found a result inconsistent with this study). According to Martens et al.

www.manaraa.com

(1997), possible explanation may be that the drivers show improved lane-keeping and reduce their speed when the lane widths decrease. Similar results to this study regarding the effect of lane width have been found by other researchers (Aguero-Valverde and Jovanis, 2009; Dong, Clarke, Richards, et al., 2014; Dong, Clarke, Yan, et al., 2014). These contradictory findings indicate that further investigation of this issue may be required.

The interaction of the area type and the LnAADT (areadt) was found to be statistically significant in both models with a negative sign. This indicates that the effects of LnAADT in urban areas are smaller than that for rural areas. This estimate actually adjusts for the effect of LnAADT on the crash frequency depending on the area type considered. On the contrary, the effect of the interaction of area type with the percentage of commercial vehicles (Areacommercial) has a positive sign, indicating that the impact of truck percentage on crash frequency is higher in urban areas. Having said that, this interaction was only found to be statistically significant in the MLE model.

Similar to the findings of Lord and Persaud (2000), the results show that not accounting for temporal correlation does not affect the way a variable affects crash frequency, but it considerably underestimates their variances. This may indicate that explanatory variables may be incorrectly attributed as significant if the temporal correlation is ignored in the model. Contrary to the results from Lord and Persaud (2000), temporal correlation affected the magnitude of the estimates in this study.

The results in Table 7 show that the interactions among area type, number of lanes, and speed limit are statistically significant. The estimates of the interaction terms in the model are graphically shown in Figure 2. From the overall trend of change in the model estimate, it can be interpreted that the increase in speed limit results in a decrease in crash frequency in urban areas (a somewhat counterintuitive result), and the change in the number of lanes does not show a consistent trend in affecting the crash frequency, except when the speed limit is low (55 mph). In that case, increasing the number of lanes decreases the crash frequency. In contrast, previous similar studies have found that an increase in the number of lanes results in an increase in crash frequency (Milton and Mannering, 1998; Abdel-Aty and Radwan, 2000; Zegeer et al., 2002; Noland and Oh, 2004; L.-Y. Chang, 2005).

www.manaraa.com

Table 7. NB model estimates (see Tables 5 and 6 for variable definitions)

Parameters	Estimation method			
	Generalized Estimating Equations		Maximum Likelihood Estimation	
	Estimate	Standard error	Estimate	Standard error
Intercept	**-18.8958**	(1.2095)	**-19.8319**	(0.7870)
lnAADT	**1.8585**	(0.0998)	**1.9049**	(0.0589)
PSRclass1	-0.0047	(0.0337)	**0.0613**	(0.0313)
PSRclass2	-0.0258	(0.0359)	-0.0077	(0.0381)
PSRclass3	0	0	0	0
Percentcommercial	-0.5613	(0.3237)	**-2.1734**	(0.3166)
Lanewidth	0.0436	(0.0593)	**0.1272**	(0.0403)
Areadt	**-1.0045**	(0.1194)	**-0.9315**	(0.0699)
Areacommercial	0.3557	(0.4614)	**2.1823**	(0.4218)
Urban_2_55	**11.4347**	(1.1854)	**10.1408**	(0.7159)
Urban_2_60	**10.9176**	(1.1930)	**9.5232**	(0.7255)
Urban_2_65	**11.0373**	(1.1907)	**9.7382**	(0.7218)
Urban_2_70	**10.7066**	(1.1661)	**9.5483**	(0.7095)
Urban_3_55	**11.3496**	(1.2071)	**10.0128**	(0.7312)
Urban_3_60	**11.2504**	(1.2104)	**9.9191**	(0.7317)
Urban_3_65	**10.9705**	(1.1992)	**9.5883**	(0.7263)
Urban_3_70	**10.6691**	(1.1851)	**9.1762**	(0.7212)
Urban_3p_55	**11.3188**	(1.2100)	**9.9131**	(0.7325)
Urban_3p_60	**11.1489**	(1.2138)	**9.7283**	(0.7342)
Urban_3p_65	**11.1138**	(1.2090)	**9.6287**	(0.7313)
Urban_3p_70	**10.7204**	(1.2325)	**9.3086**	(0.7424)
QIC	-152184.0149		N/A	
QICu	-152184.6453		N/A	

Bold estimates are significant at 95% level of confidence (p-value < 0.05)

www.manaraa.com

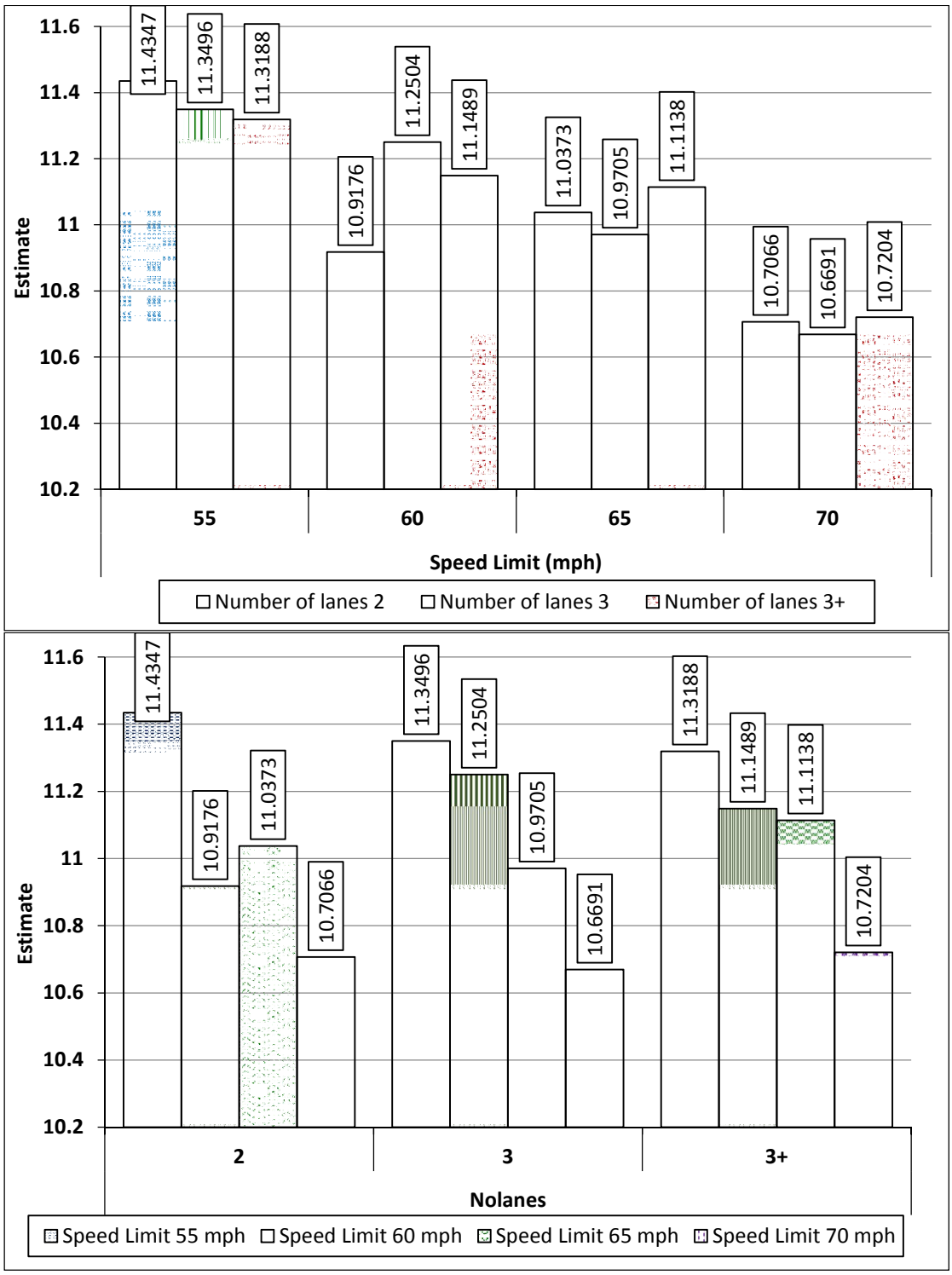

Figure 2. Estimates of the interaction terms between number of lanes, and speed limit in urban areas (Top: effect of number of lanes within each class of speed limit, Bottom: effect of speed limit within each class of number of lanes)

www.manaraa.com

The significance of these effects on crash frequency, however, cannot be directly determined from Table 7, because the p-values which determine the statistical significance are tailored to determining whether these categories have significantly different effects relative to the base category, which in this case is rural_2_70. While the effect of moving from one category to another can be computed by subtracting the slope estimates of one category from the other one, the statistical significance of this difference had to be tested using an additional step. For example, suppose we wish to determine the effect of changing the road segment from 2 to 3 lanes with the speed limit held constant at 60 mph. This effect is the difference between the slopes of urban_3_60 and urban_2_60, which is 11.2504 - 10.9176 = 0.3328. To determine if this difference is statistically significant, we redefined the dummy variable urban_2_60 to take the value of 1 when SL=60, nolanes=2, or when SL=60 and nolanes=3. This new dummy variable would be the appropriate variable to use in the model if there is no difference between the (SL=60, nolanes=2) category and the (SL=60, nolanes=3) category. This is the null hypothesis for the test we are conducting. This dummy variable was fitted in the model along with urban_3_60. If the estimate for urban_3_60 is found to be statistically significant, then that means that there is significant deviation from the null hypothesis and the change from nolanes=2 to nolanes=3 for urban road segments with speed limit of 60 mph, is statistically significant. Tables 8 and 9 show the results for the GEE model for all the possible changes that can be made between the number of lanes and the speed limit, using the GEE and MLE methodologies, respectively.

Studying the results of the analysis (Tables 8 and 9) reveals that the standard errors estimated using MLE method are higher than those estimated by the GEE method, as was also the case for the other main factors of the model. This results from not accounting for the temporal correlation between the yearly observations of the same segments over the years. These results are consistent with others in the literature (Mannering and Bhat, 2014). The shaded areas in both tables show the statistically significant effects of the change in the number of lanes or speed limits. For example, changing the number of lanes from 2 to 3 lanes while keeping the speed limit of 70 mph, has a negative effect of $e^{-0.372}$ on crash frequency in the MLE model, which is statistically

www.manaraa.com

significant compared to the same change examined using the GEE method with an effect of $e^{-0.038}$, though it is not statistically significant.

Table 8. Analysis of statistical significance of the effect of change in the number of lanes and speed limit on crash frequency (using generalized estimating equation method)

NoLanes		2				3				3+			
	SL	55	60	65	70	55	60	65	70	55	60	65	70
2	55	Est.	**-0.517**	**-0.397**	**-0.728**	-0.085	-0.184	**-0.464**	**-0.766**	-0.116	**-0.286**	-0.321	**-0.714**
		SD	0.178	0.154	0.143	0.143	0.156	0.140	0.202	0.159	0.145	0.172	0.332
	60	**0.517**	Est.	0.120	-0.211	**0.432**	**0.333**	0.053	-0.249	**0.401**	0.231	0.196	-0.197
		0.178	SD	0.168	0.154	0.165	0.168	0.158	0.216	0.182	0.163	0.189	0.341
	65			Est.	**-0.331**	**0.312**	0.213	-0.067	-0.368	0.282	0.112	0.077	-0.317
				SD	0.136	0.137	0.146	0.120	0.197	0.156	0.134	0.161	0.329
	70				Est.	**0.643**	**0.544**	**0.264**	-0.038	**0.612**	**0.442**	**0.407**	0.014
					SD	0.141	0.151	0.127	0.185	0.160	0.142	0.166	0.328
3	55					Est.	-0.099	**-0.379**	**-0.681**	-0.031	-0.201	-0.236	-0.629
						SD	0.126	0.111	0.193	0.135	0.111	0.150	0.323
	60						Est.	**-0.280**	**-0.581**	0.068	-0.102	-0.137	-0.530
							SD	0.122	0.200	0.148	0.114	0.157	0.327
	65							Est.	-0.301	**0.348**	0.178	0.143	-0.250
								SD	0.186	0.136	0.108	0.139	0.321
	70								Est.	**0.650**	**0.480**	**0.445**	0.051
									SD	0.208	0.193	0.214	0.348
3+	55									Est.	-0.170	-0.205	-0.598
										SD	0.133	0.158	0.332
	60										Est.	-0.035	-0.429
											SD	0.142	0.321
	65											Est.	-0.393
												SD	0.336
	70											0.393	Est.
												0.336	SD

Note. Nolanes, and SL represent the number of lanes and speed limit in mph, respectively. Est. presents the model estimates (effect of change in nolanes or SL), and SD stands for estimate's standard error. Bold values are statistically significant at 95% level of confidence.

Figure 3 presents a graphical comparison of the standard errors of the models' estimates. It can be observed that the standard errors for the GEE model estimates are higher than those for their MLE model counterparts, except for the variable PSR, which has a subtle difference in standard error values. This indicates that the MLE model ignores serial correlation, underestimates the variance of the coefficient estimates, resulting in more significant factors. Some explanatory variables may become insignificant when temporal correlation is considered (Lord and Persaud, 2000), which is also the case here.

www.manaraa.com

Table 9. Statistical significance of the effect of change in the number of lanes and speed limit in the model estimated by the method of maximum likelihood estimation

NoLanes		2				3				3+			
	SL	55	60	65	70	55	60	65	70	55	60	65	70
2	55	Est.	**-0.618**	**-0.403**	**-0.593**	-0.128	**-0.222**	**-0.553**	**-0.965**	**-0.228**	**-0.413**	**-0.512**	**-0.832**
		SD	0.105	0.090	0.081	0.086	0.090	0.082	0.110	0.099	0.083	0.105	0.172
	60	**0.618**	Est.	**0.215**	0.025	**0.490**	**0.396**	0.065	**-0.347**	**0.390**	**0.205**	0.106	-0.215
		0.105	SD	0.105	0.091	0.102	0.106	0.098	0.124	0.114	0.099	0.119	0.182
	65			Est.	**-0.190**	**0.275**	**0.181**	-0.150	**-0.562**	0.175	-0.010	-0.110	**-0.430**
				SD	0.083	0.084	0.088	0.080	0.110	0.097	0.080	0.104	0.172
	70				Est.	**0.465**	**0.371**	0.040	**-0.372**	**0.365**	**0.180**	0.081	-0.240
					SD	0.085	0.090	0.078	0.105	0.100	0.083	0.104	0.172
3	55					Est.	-0.094	**-0.425**	**-0.837**	-0.100	**-0.285**	**-0.384**	**-0.704**
						SD	0.075	0.069	0.106	0.086	0.063	0.094	0.166
	60						Est.	**-0.331**	**-0.743**	-0.006	**-0.191**	**-0.290**	**-0.611**
							SD	0.074	0.109	0.090	0.069	0.098	0.168
	65							Est.	**-0.412**	**0.325**	**0.140**	0.041	-0.280
								SD	0.102	0.085	0.062	0.093	0.165
	70								Est.	**0.737**	**0.552**	**0.453**	0.132
									SD	0.117	0.103	0.122	0.183
3+	55									Est.	**-0.185**	**-0.284**	**-0.605**
										SD	0.080	0.106	0.173
	60										Est.	-0.100	**-0.420**
											SD	0.090	0.163
	65											Est.	-0.320
												SD	0.177
	70											0.320	Est.
												0.177	SD

Note. Nolanes, and SL represent the number of lanes and speed limit in mph, respectively. Est. presents the model estimates (effect of change in nolanes or SL), and SD stands for estimate's standard error. Bold values are statistically significant at 95% level of confidence.

Figure 4 presents the results of the χ^2-values for the variables used in the models. The χ^2-values for the PSRclass1, PSRclass2, percentcommercial, lanewidth, and areacommercial were too small to be visible in the figure. Their values are shown in Table 10. By comparing the χ^2-values of the GEE model with the MLE model, it can be observed that almost all of the χ^2-values for the GEE model are lower than those for the MLE model. This is also an indication that the GEE model incorporates temporal correlation and provides more reliable estimates compared to the MLE model. The decrease in χ^2-value results in a higher p-value, and consequently makes the variables PSRclass1, percentcommercial, lanewidth, and the interaction variable areacommercial

www.manaraa.com

insignificant at 95% level of confidence. A lower χ^2-value indicates a better fit of the model (Allison, 2012). The only statistic that might make a difference in the resulting conclusion is the χ^2-value of the PSRclass2, which is larger for the GEE model. Furthermore, this result verifies that the autoregressive correlation structure is an appropriate form of correlation to be used for this type of data.

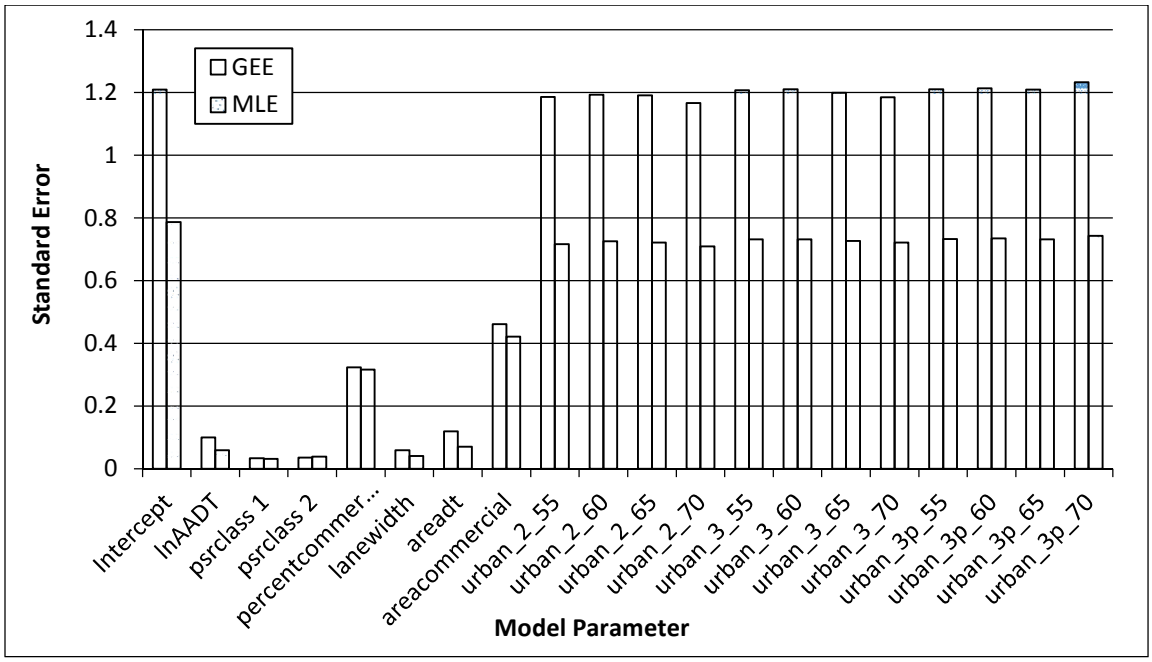

Figure 3. Comparison of the models' standard errors using generalized estimating equations and maximum likelihood estimation methods (see Tables 5 and 6 for variable definitions)

Lin et al. (2002) have discussed the cumulative residual (CURE) method to investigate models' quality of fit. This method generates a plot in which the cumulative residuals are plotted for an independent variable of the model and compared against the zero-residual line (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006; Lord and Park, 2008). Lord and Persaud (2000) found that the crash models that incorporate time trend usually perform better than traditional models without time trend. In this study, CURE method was also used to evaluate the goodness of fit. This was done using the ASSESS option of the GENMOD procedure in the SAS code written for the model (SAS, 2008). Figure 5 shows an example of the CURE plots for the independent variable LnAADT, generated by SAS, for the GEE model with AR(1) correlation structure and MLE model.

www.manaraa.com

Table 10. Comparison of relatively smaller χ^2-values (see Tables 5 and 6 for variable definitions)

Parameter	Estimation method	
	Generalized estimating equation	Maximum likelihood estimation
PSRclass1	0.0196	3.84
PSRclass2	0.5184	0.04
Percentcommercial	2.9929	47.12
Lanewidth	0.5476	9.97
Areacommercial	0.5929	26.76

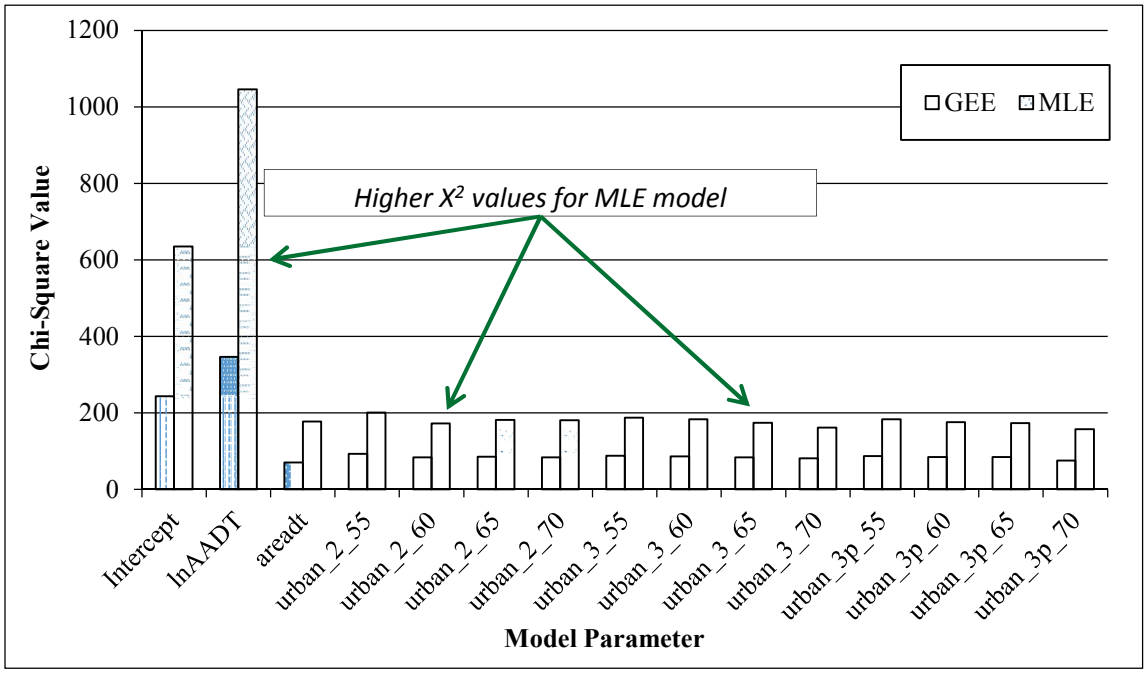

Figure 4. Comparison of the models' χ^2-values using generalized estimating equations and maximum likelihood estimation methods (see Tables 5 and 6 for variable definitions)

The graph presents the actual cumulative residuals for the model (bold line) and the simulated residual paths (dotted lines). In order to evaluate an entire model these cumulative residual graphs were produced for all the variables and the link function. A comparison of the CURE plots for the independent variables in the GEE model and the MLE model indicated that the actual residual pattern for the GEE model is closer to the expected patterns generated by simulation. Also, similar to the result of the study conducted by Wang and Abdel-Aty (2006), in this study, higher p-values for the CURE test were obtained for the GEE model compared to the MLE model. A comparison of

www.manaraa.com

these CURE plots in this study indicated a similar result to the study by Lord and Persaud (2000), confirming that the GEE model with temporal correlation is an improved crash frequency model with less biased and more accurate coefficient estimates.

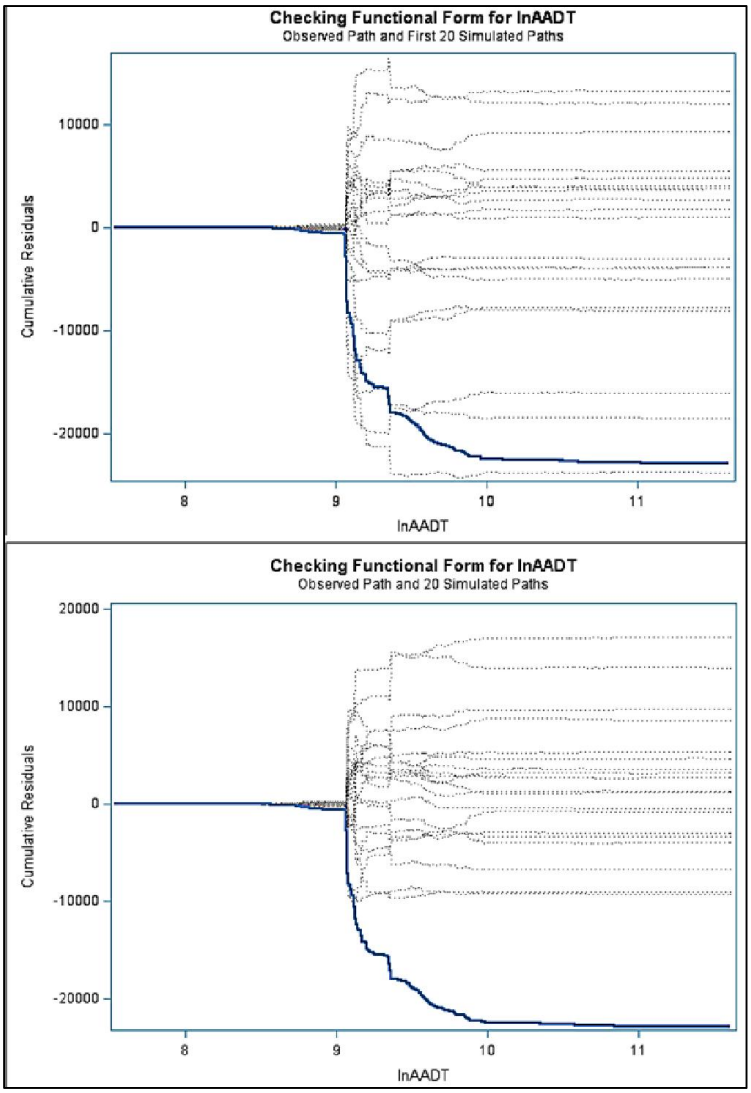

Figure 5. Cumulative residuals plot for LnAADT for the negative binomial models estimated using the methods of generalized estimating equation (top) and maximum likelihood estimation (bottom)

4.2. Validation of correlation structure

As a validation on the possible effects of unobserved hetrogeneities on the investigated temporal serial correlation, the estimates of the GEE models with three, seven, and 10 years of analysis periods (the same periods as were used for examining the correlation values) were compared to verify the assumption of having no unobserved

www.manaraa.com

hetrogeneity left unaccounted for in the longitudinal crash data. Table 11 shows the results of the GEE model with AR(1) correlation structure for the three above-mentioned analysis periods. It can be observed that the covariate estimates and their significance in the model are very similar to each other. This result verifies the correctness of the assumption that AR correlation structure is present in the longitudinal crash data and further endorses the outcomes of the study.

Table 11. Negative binomial model estimates using generalized estimating equations for three different analysis periods (see Tables 5 and 6 for variable definitions)

Parameter	3 years (2002-2004)		7 years (2002-2008)		10 years (2002-2011)	
	Estimate	Sd. Error	Estimate	Sd. Error	Estimate	Sd. Error
Intercept	**-19.1238**	1.5903	**-18.5554**	1.2734	**-18.8958**	1.2095
lnAADT	**1.9011**	0.143	**1.8462**	0.1092	**1.8585**	0.0998
PSRclass1	0.0315	0.0535	0.0318	0.0374	-0.0047	0.0337
PSRclass2	-0.016	0.0545	0.0151	0.0369	-0.0258	0.0359
PSRclass3	0	0	0	0	0	0
Percentcommercial	-0.3908	0.4322	-0.5687	0.3424	-0.5613	0.3237
Lanewidth	0.0327	0.0641	0.0313	0.0576	0.0436	0.0593
Areadt	**-0.979**	0.1745	**-1.012**	0.1297	**-1.0045**	0.1194
Areacommercial	0.3306	0.6253	0.1616	0.4875	0.3557	0.4614
Urban_2_55	**11.1303**	1.7393	**11.4822**	1.2912	**11.4347**	1.1854
Urban_2_60	**10.7334**	1.7334	**10.9673**	1.298	**10.9176**	1.193
Urban_2_65	**10.7175**	1.7323	**11.0483**	1.2953	**11.0373**	1.1907
Urban_2_70	**10.5082**	1.6966	**10.747**	1.2711	**10.7066**	1.1661
Urban_3_55	**11.0502**	1.759	**11.3884**	1.3127	**11.3496**	1.2071
Urban_3_60	**10.9185**	1.7637	**11.251**	1.3155	**11.2504**	1.2104
Urban_3_65	**10.6881**	1.7503	**11.0405**	1.3053	**10.9705**	1.1992
Urban_3_70	**10.1052**	1.7329	**10.793**	1.2934	**10.6691**	1.1851
Urban_3p_55	**11.0791**	1.7652	**11.3817**	1.316	**11.3188**	1.21
Urban_3p_60	**10.7896**	1.7697	**11.2124**	1.3194	**11.1489**	1.2138
Urban_3p_65	**10.8692**	1.762	**11.1962**	1.3147	**11.1138**	1.209
Urban_3p_70	**10.3051**	1.7811	**10.7672**	1.3418	**10.7204**	1.2325

* Bold values are significant at 95% level of confidence

www.manaraa.com

5. CONCLUSIONS

The objective of this study was to use the generalized estimating equations (GEE) method to develop a longitudinal negative binomial (NB) model for analysis of the interstate highways of Missouri over the years 2002 through 2011. General modeling approaches used in other research studies usually neglect to account for the temporal correlation in crash frequencies observed over several years. The GEE procedure overcomes these difficulties in developing unbiased estimates by accommodating temporal correlation in crash observations and not underestimating the variation in coefficient estimates.

A GEE model was developed using autoregressive type 1 correlation structure and compared to an equivalent MLE estimation model that do not account for the temporal correlation. This study examined the standard errors and the Chi-square values of the variables estimated using GEE and MLE methods along with evaluation of the cumulative residual plots for the two models. The GEE model, allowing for temporal correlations proved to be a superior model compared to the traditional NB model using MLE method, providing more accurate and less biased model estimates. This result is in agreement with the literature (Lord and Persaud, 2000; Ulfarsson and Shankar, 2003; Mannering and Bhat, 2014).

The natural logarithm of AADT (*LnAADT*) was found to be a statistically significant factor with a positive sign in both models (GEE and MLE), indicating higher number of crashes with higher traffic volume. Also, the significance of the interaction of the *LnAADT* with the *area type* with a negative estimate in both models showed that the effect of traffic volume in urban areas is smaller than rural areas. An increase in speed limit was found to result in a decrease in crash frequency in urban areas (a somewhat counterintuitive result), and the change in the number of lanes did not show a consistent trend in affecting the crash frequency, except when the speed limit was low at 55 mph. In that case, increasing the number of lanes results in a decrease in the crash frequency which is in contrast to the results of some of the previous studies (Noland and Oh, 2004; L.-Y. Chang, 2005).

www.manaraa.com

By considering temporal correlation in the model, some explanatory variables may become insignificant which was also the case found in this study. *Percent commercial* with a negative estimate was only found to be significant in the MLE model indicating that heavy vehicles result in fewer crashes. This indicates drivers may use more caution and reduce their speed when they travel close to large vehicles. Carson and Mannering (2001) and Lao et al. (2014) also found similar results of the effect of truck percentage on ice-related crash and rear-end crash occurrences, respectively. Also, the interaction of this variable with the *area type* was found to be significant with a positive sign indicating that the impact of commercial vehicles on crash frequency is higher in urban areas. Another factor that was only found to be significant in the MLE model was *lane width*. The positive sign of the estimate for lanewidth, however, seems to be counterintuitive (For example, see the works of Li et al. (2008), Manuel et al. (2014), whose results are inconsistent with this study); however, some other recent studies have found similar results to this study regarding the effect of lane width (Aguero-Valverde and Jovanis, 2009; Dong, Clarke, Richards, et al., 2014; Dong, Clarke, Yan, et al., 2014). Martens et al. (1997) note that when the lane widths decrease drivers show improved lane-keeping and reduce their speed. These inconsistent conclusions indicate that further study of this matter may be required.

Furthermore, the autoregressive correlation structure was found to be an appropriate structure for this longitudinal type of data. If crash data is available for several years, it is recommended to use larger data sets to increase the model reliability, but also to incorporate temporal correlations when modeling crashes are aggregated over several years. This provides more accurate crash frequency models and therefore, safety policies and crash countermeasures based on such models will be more efficient in saving lives and resources. This study confirms that the use of GEE is a good approach for addressing the serial correlation in crash frequency data.

6. ACKNOWLEDGEMENTS

The authors of this paper would like express their appreciation to the reviewers of the *Analytic Methods in Accident Research*, who helped improve the quality of this paper.

www.manaraa.com

7. REFERENCES

Abdel-Aty, M., and Radwan, A. E. 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention, 32*(5), 633-642.

Aguero-Valverde, J., and Jovanis, P. P. 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record: Journal of the Transportation Research Board, 2136*, 82-91.

Allison, P. D. 2012. *Logistic regression using SAS: Theory and application*: SAS Institute.

ASTM-D6433-07. 2007. Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys. West Conshohocken, PA: ASTM International.

ASTM-E1489-08. 2008. Standard Practice for Computing Ride Number of Roads from Longitudinal Profile Measurements Made by an Inertial Profile Measuring Device. West Conshohocken, PA: ASTM International.

Ballinger, G. A. 2004. Using generalized estimating equations for longitudinal data analysis. *Organizational research methods, 7*(2), 127-150.

Belsley, D. A., Kuh, E., and Welsch, R. E. 2005. *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571): John Wiley and Sons.

Bhat, C. R., Born, K., Sidharthan, R., and Bhat, P. C. 2014. A count data model with endogenous covariates: Formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research, 1*, 53-71.

Carson, J., and Mannering, F. 2001. The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis and Prevention, 33*(1), 99-109.

Castro, M., Paleti, R., and Bhat, C. R. 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B, 46*(1), 253-272.

Chang, H. L., Woo, T. H., and Tseng, C. M. 2006. Is rigorous punishment effective? A case study of lifetime license revocation in Taiwan. *Accident Analysis and Prevention, 38*(2), 269-276.

www.manaraa.com

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

Chi, G., McClure, T. E., and Brown, D. B. 2012. Gasoline prices and traffic crashes in Alabama, 1999–2009. *Traffic Injury Prevention, 13*(5), 476-484.

Cox, D. R. 1984. Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1-24.

Dong, C., Clarke, D. B., Richards, S. H., and Huang, B. 2014. Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: An exploratory analysis. *Accident Analysis and Prevention, 62*, 87-94.

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

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

Dupont, E., Papadimitriou, E., Martensen, H., and Yannis, G. 2013. Multilevel analysis in road safety research. *Accident Analysis and Prevention*.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. 1993. Regression models for discrete longitudinal responses. *Statistical Science*, 284-299.

Gill, J. 2001. *Generalized linear models: a unified approach* (Vol. 134): Sage Publications, Incorporated.

Giuffrè, O., Granà, A., Giuffrè, T., and Marino, R. 2007. Improving reliability of road safety estimates based on high correlated accident counts. *Transportation Research Record: Journal of the Transportation Research Board, 2019*, 197-204.

Giuffrè, O., Grana, A., Giuffrè, T., and Marino, R. 2013. Accounting for Dispersion and Correlation in Estimating Safety Performance Functions. An Overview Starting from a Case Study. *Modern Applied Science, 7*(2), p11.

Guo, F., Wang, X., and Abdel-Aty, M. 2010. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis and Prevention, 42*(1), 84-92.

www.manaraa.com

Hanley, J. A., Negassa, A., and Forrester, J. E. 2003. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology, 157*(4), 364-375.

Hardin, J. W., and Hilbe, J. M. (2007). Generalized Estimating Equations *Wiley Encyclopedia of Clinical Trials*: John Wiley and Sons, Inc.

Hauer, E., and Bamfo, J. 1997. *Two tools for finding what function links the dependent variable to the explanatory variables.* Paper presented at the Proceedings of the ICTCT 1997 Conference.

Hutchings, C. B., Knight, S., and Reading, J. C. 2003. The use of generalized estimating equations in the analysis of motor vehicle crash data. *Accident Analysis and Prevention, 35*(1), 3-8.

Kutner, M. H., Nachtsheim, C., and Neter, J. 2004. Applied linear regression models.

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

Lenguerrand, E., Martin, J. L., and Laumon, B. 2006. Modelling the hierarchical structure of road crash data—Application to severity analysis. *Accident Analysis and Prevention, 38*(1), 43-53.

Li, X., Lord, D., Zhang, Y., and Xie, Y. 2008. Predicting motor vehicle crashes using support vector machine models. *Accident Analysis and Prevention, 40*(4), 1611-1618.

Liang, K.-Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Lin, D., Wei, L., and Ying, Z. 2002. Model-Checking Techniques Based on Cumulative Residuals. *Biometrics, 58*(1), 1-12.

Littell, R. C., Stroup, W. W., and Freund, R. J. 2002. *SAS for linear models*: SAS Institute.

Lord, D., and Persaud, B. N. 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record: Journal of the Transportation Research Board, 1717*, 102-108.

www.manaraa.com

Lord, D., and Park, P. Y.-J. 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis and Prevention, 40*(4), 1441-1457. doi: http://dx.doi.org/10.1016/j.aap.2008.03.014

Lord, D., and Mahlawat, M. 2009. Examining Application of Aggregated and Disaggregated Poisson-Gamma Models Subjected to Low Sample Mean Bias. *Transportation Research Record: Journal of the Transportation Research Board, 2136*, 1-10.

Maas, C. J., and Hox, J. J. 1999. Sample sizes for multilevel modeling. *Am J Public Health, 89*, 1181-1186.

Maher, M. J., and Summersgill, I. 1996. A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis and Prevention, 28*(3), 281-296.

Mancl, L. A., and DeRouen, T. A. 2001. A covariance estimator for GEE with improved small-sample properties. *Biometrics, 57*(1), 126-134.

Mannering, F. L., and Bhat, C. R. 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research, 1*, 1-22.

Manuel, A., El-Basyouny, K., and Islam, M. T. 2014. Investigating the safety effects of road width on urban collector roadways. *Safety science, 62*, 305-311.

Martens, M., Compte, S., and Kaptein, N. A. 1997. The effects of road design on speed behaviour: a literature review.

McCullagh, P., and Nelder, J. A. 1989. *Generalized linear model* (Vol. 37): Chapman and Hall/CRC.

Méndez, Á. G., Aparicio Izquierdo, F., and Ramírez, B. A. 2010. Evolution of the crashworthiness and aggressivity of the Spanish car fleet. *Accident Analysis and Prevention, 42*(6), 1621-1631.

Milton, J. C., and Mannering, F. L. 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation, 25*(4), 395-413.

www.manaraa.com

Mohammadi, M. A., Samaranayake, V. A., and Bham, G. 2013. *The Effect of Incorporating Temporal Correlations into Negative Binomial Count Data Models.* Paper presented at the Fourth International Conference on Road Safety and Simulation, Rome, Italy.

Mohammadi, M. A. (2014). *Longitudinal analysis of crash frequency data.* Doctoral dissertation, Missouri University of Sicence and Technology.

Mohammadi, M. A., Samaranayake, V. A., and Bham, G. 2014. *Safety Effect of Missouri's Strategic Highway Safety Plan - Missouri's Blueprint for Safer Roadways.* Paper presented at the Transportation Research Board 93rd Annual Meeting., Washington. D.C.

Nelder, J. A. 1977. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, 48-77.

Noland, R. B., and Oh, L. 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. *Accident Analysis and Prevention, 36*(4), 525-532.

Noland, R. B., Quddus, M. A., and Ochieng, W. Y. 2008. The effect of the London congestion charge on road casualties: an intervention analysis. *Transportation, 35*(1), 73-91.

Pan, W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics, 57*(1), 120-125.

Park, B.-J., and Lord, D. 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention, 41*(4), 683-691. doi: http://dx.doi.org/10.1016/j.aap.2009.03.007

Peng, Y., Boyle, L. N., and Hallmark, S. L. 2012. Driver's lane keeping ability with eyes off road: Insights from a naturalistic study. *Accident Analysis and Prevention*.

Poch, M., and Mannering, F. 1996. Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering, 122*(2), 105-113.

Quddus, M. A. 2008. Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention, 40*(5), 1732-1741.

SAS. 2008. *SAS/STAT 9.2 User's Guide: The GENMOD Procedure (book Excerpt)*: SAS Institute.

www.manaraa.com

Savolainen, P. T., and Tarko, A. P. 2005. Safety impacts at intersections on curved segments. *Transportation Research Record: Journal of the Transportation Research Board, 1908*, 130-140.

Shankar, V. N., Mannering, F., and Barfield, W. 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention, 27*(3), 371-389.

Stavrinos, D., Jones, J. L., Garner, A. A., Griffin, R., Franklin, C. A., Ball, D., Welburn, S. C., Ball, K. K., Sisiopiku, V. P., and Fine, P. R. 2013. Impact of distracted driving on safety and traffic flow. *Accident Analysis and Prevention*.

Ulfarsson, G. F., and Shankar, V. N. 2003. Accident count model based on multiyear cross-sectional roadway data with serial correlation. *Transportation Research Record: Journal of the Transportation Research Board, 1840*, 193-197.

Venkataraman, N., Ulfarsson, G. F., Shankar, V., Oh, J., and Park, M. 2011. Model of Relationship Between Interstate Crash Occurrence and Geometrics. *Transportation Research Record: Journal of the Transportation Research Board, 2236*, 41-48.

Venkataraman, N., Ulfarsson, G. F., and Shankar, V. N. 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis and Prevention, 59*, 309-318.

Venkataraman, N., Shankar, V., Ulfarsson, G. F., and Deptuch, D. 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. *Analytic Methods in Accident Research, 2*, 12-20.

Wang, X., and Abdel-Aty, M. 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accident Analysis and Prevention, 38*(6), 1137-1150.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. 2011. *Statistical and econometric methods for transportation data analysis*: CRC press.

Xiong, Y., Tobias, J. L., and Mannering, F. L. 2014. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. *Transportation Research Part B, 67*, 109-128.

www.manaraa.com

Zegeer, C. V., Stewart, J. R., Huang, H. H., and Lagerwey, P. A. 2002. Safety effects of marked vs. unmarked crosswalks at uncontrolled locations: Executive summary and recommended guidelines.

Zeger, S. L., and Liang, K.-Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121-130.

Zorn, C. J. 2001. Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 470-490.

Zou, Y., Zhang, Y., and Lord, D. 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. *Analytic Methods in Accident Research, 1*, 39-52.

www.manaraa.com

III. SEASONAL EFFECTS OF CRASH CONTRIBUTING FACTORS ON HIGHWAY SAFETY

ABSTRACT

A longitudinal negative binomial model is developed in this paper that takes into account the seasonal effects of crash causality factors based on ten years (2002-2011) of Missouri Interstate highway crash data. The technique of generalized estimating equation (GEE) with autoregressive correlation structure is used. The results explain the overall effect of seasonality and whether the magnitude and/or type of various effects are different according to climatic changes. Traffic volume was found to have an appreciable effect in increasing the crash occurrence in spring and lower effect in winter, compared to the fall season. Fewer crashes were associated with higher pavement serviceability (measure of pavement surface quality, higher value is better) and this effect was found to be highest in the spring season followed by summer and winter, again when compared to the fall season. Heavy vehicles were found to reduce the likelihood of crash occurrences and this effect is higher in urban areas; although compared to other times of the year, the effect of heavy vehicles is lower during the summer season. The results indicated that the fall season is associated with the lowest crash frequency compared to the other seasons; winter season having the highest impact followed by summer and spring. This paper also evaluated the effects of the Missouri's Strategic Highway Safety Plan (MSHSP) implemented from 2005-2011. The plan was found to be effective as it reduced the crash frequency. Similar strategic plans therefore should be initiated in the future as well.

Keywords: seasonal effect, seasonality, generalized estimation equation, crash frequency model, strategic highway safety plan

1. INTRODUCTION

Traffic safety in transportation networks is one of the main priorities for many government agencies, private organizations and the society as a whole. This is mainly

www.manaraa.com

due to the significant monetary and non-monetary costs associated with crashes (Elvik, 2000). The National Highway Traffic Safety Administration (NHTSA) estimated the economic cost of motor vehicle crashes to be around $230 billion a year in 2000 (NHTSA, 2008). In 2009, an estimated 5,505,000 traffic crashes occurred, which included 33,808 fatalities and 2,217,000 people were injured (NHTSA, 2009). Peden et al. (2004) have forecasted that road related injuries are expected to increase from ranked ninth in 1990 to the third largest contributor to the global burden of disease and injury in 2020. This immense loss of life and property to society from motor vehicle crashes warrants accurate identification of crash contributing factors and countermeasures.

Crash frequency is a fundamental indicator of "safety" in terms of evaluation and estimation (HSM, 2010). The term "crash evaluation" refers to determining the effectiveness of a particular treatment after its implementation. Safety investigators have continually sought ways to gain a better understanding of crash causes and propose measures to reduce it (Lord and Mannering, 2010). Crashes mainly occur due to factors that stem from drivers' behavior, vehicular characteristics, highway design, and environmental conditions. Geographical location and climatic environment, particularly weather can be a major factor that contributes to the occurrence of crashes (Garber and Hoel, 2008a). Few studies in the crash evaluation realm deal with the seasonal effects of crashes that incorporate crash causality factors. Hilton et al. (2011) provided a basic understanding of the seasonal patterns of fatal crashes particularly in the summer and showed that with a better understanding of the crash causes over different times of the year, policy-makers can improve the safety of specific roadway segments according to the climatic conditions. Ahmed et al. (2011) and Yu et al. (2013a) have demonstrated a significant seasonal effect on crash frequencies in mountainous terrain with adverse weather conditions and that different traffic management strategies should be designed based on seasons. Yang et al. (2013) accounted for the seasonal fluctuations of crash frequency in work zones by adjusting the AADT using seasonal correction factors. A survey conducted by the Center For Excellence in Rural Safety on seasonal crash frequency perceptions identified that the general public (83% of participants) believed that winter is the most dangerous time to drive due to hazardous road conditions during and after a snow fall (CERS, 2010).

www.manaraa.com

To the best of the authors' knowledge, there are no in-depth analyses of seasonality effects of crash causes. The objective of this study is to investigate the seasonal effects on crash causality by developing a longitudinal negative binomial model using several years of crash data on Interstate highways of Missouri. The results of this paper will help in developing policies regarding highway safety countermeasures with insight on the effects of seasonal changes on roadway crashes. For example, a previous study conducted by Carson and Mannering (2001) found no significant effects on the use of warning signs in reducing the ice related crash frequency or severity. They believed that the maintenance cost and personnel for those signs required a justification. This paper uses generalized estimating equation technique to develop a negative binomial crash frequency model with longitudinal crash data to address the seasonal effects on crash causal factors and the interventions of the Missouri Strategic Highway Safety Plan (MSHSP).

2. METHODOLOGY

The technique of generalized estimating equations (GEE) is used in this paper for correlated crash data (as a result of repeated observations over time) to estimate the model parameters. Zeger and Liang (1986) first used the technique of GEE by using generalized linear models for repeated observations. Since then, many research studies have used this methodology to account for the temporal correlation amongst the observations made from the same unit of analysis (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006; Giuffrè et al., 2007). Consider a model of crash frequency observations at a highway segment i during time t (Y_{it}) and k covariates (X_{it}). According to Zorn (2001) the relationship between Y_i and X_i can be shown as:

$$\mu_i = E(Y_i) = h(X_i\beta) \tag{1}$$

where,

μ_i : Expected value of the crash frequency at segment i, (Y_i), $i = 1, 2, \ldots, N$

β : $k \times 1$ vector of estimable parameters

X_i : $t \times k$ matrix of covariates for segment i $(i = 1, 2, \ldots N, t = 1, 2, \ldots T)$.

The variance of Y_i, (V_i) is specified as a function of the mean:

www.manaraa.com

$$V_i = g(\mu_i)/\phi \tag{2}$$

Where, ϕ is the scale parameter. The quasi-likelihood estimate of β is then the solution to a set of k "quasi-score" differential equations (Zeger and Liang, 1986; Zorn, 2001):

$$U_k(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1}(Y_i - \mu_i) = 0 \tag{3}$$

where,

$D_i = \mu_i/\beta,$

$V_i = \dfrac{(A_i)^{1/2} R_i(\alpha)(A_i)^{1/2}}{\phi}$

A_i : $T \times T$ diagonal matrices with $g(\mu_{it})$ as the t^{th} diagonal element,

$R_i(\alpha)$: a $T \times T$ matrix of the working correlations across time for a given Y_i , and

α : vector of unknown parameters with a specific structure (according to the type of correlation structure).

Substituting Equation (4) into Equation (3) results in the GEE estimators and it can be seen that it reduces to a generalized linear model when $T = 1$ (Zorn, 2001). Every element of the correlation matrix R_i should be known in order to solve the GEE; however, the exact correlation type for the repeated measurements is not always known. An alternative approach suggested by (Zeger and Liang, 1986) is to use a "working" matrix $\hat{V}$ of the correlation matrix V_i, based on the correlation matrix $\widehat{R}_i$, which results in estimating the β parameters using the following differential equations:

$$U_k(\beta) = \sum_{i=1}^{N} D_i' \hat{V}_i^{-1}(Y_i - \mu_i) = 0 \tag{4}$$

The covariance matrix of Equation 5 is given by

$$cov(\hat{\beta}) = \sigma^2 \left[\sum_{i=1}^{N} D_i' \hat{V}_i^{-1} D_i\right]^{-1} \left[\sum_{i=1}^{N} D_i' \hat{V}_i^{-1} V_i \hat{V}_i^{-1} D_i\right] \left[\sum_{i=1}^{N} D_i' \hat{V}_i^{-1} D_i\right]^{-1} \tag{5}$$

Using this methodology, $\hat{\beta}$ provides consistent estimates of β even if the correlation matrix V_i is estimated inadequately and the confidence interval for β is correct. Therefore, the need to know the type of correlation is eliminated even when the covariance matrix is specified incorrectly. However, it has been argued that there should

المنارة للاستشارات
www.manaraa.com

be no missing observations for any segment to assume that $\hat{\beta}$ is a correct estimate of β, otherwise, coefficient estimates will be biased (Lord and Persaud, 2000).

The potential positive autocorrelation in the Missouri crash data was examined by the Durbin-Watson (DW) test. The results indicated the presence of a positive autocorrelation. More details of this test is presented in a study by Mojtaba Ale Mohammadi et al. (2014b). In the data used for this study, all the covariates were considered at the segment level and it is assumed that there are no unobserved heterogeneity effects on the considered covariates of the model. Mojtaba Ale Mohammadi et al. (2014b) have shown that this is a valid assumption for the data used in this study.

3. CRASH DATA AND MODEL VARIABLES

Missouri Department of Transportation (DOT) provided the data used in this study. Ten years (2002-2011) of crash data used consisted of all crashes that occurred over the following Interstate highways: I-29, I-35, I-44, I-49, I-55, and I-70 with an overall length of 1169 miles. The highways were more than 100 miles long and covered different parts of the state. A total number of 7,742 unique segments with an average length of 2.29 miles were identified. Missouri DOT determined the boundaries of these segments based on the homogeneity of the geometric and traffic properties. The total number of crashes analyzed in this study was 126,211, 64.7% of which occurred in urban areas. For the four seasons, the minimum and maximum number of crashes analyzed was 2523 and 4039, respectively.

The explanatory variables selected for the analysis were the area type (urban or rural), number of lanes (range 2 to 7 lanes), lane width (min of 10 ft. to max of 18 ft.), AADT (min of 4198 to max of 101594 vehicles per day), speed limit (min of 55 to max of 70 mph), PSR (pavement serviceability rate, ranging from 17.4 to 66.4), and truck percentage (3% to 67%). A higher value of PSR indicates a healthier pavement condition. The high truck percentages observed are located on transit highways with low traffic and night time trucks. A similar data set was utilized in a study by Mojtaba Ale Mohammadi et al. (2014b). Additionally, four dummy variables were created to account for the

www.manaraa.com

seasonal variations: Spring_dummy, Summer_dummy, Fall_dummy, and Winter_dummy. The Fall_dummy variable was used as the base category against which the effects of the other seasons were statistically examined. Table 1 presents the list of continuous and dummy variables used in the analysis.

Multicollinearity among variables was checked, but no variables were found. This was conducted to reduce any inflation of the standard errors and to stabilize the estimated effects of the variables. Specific details of the multicollinearity check can be found elsewhere (Mojtaba Ale Mohammadi et al., 2014b). In order to avoid the confounding effects observed amongst the variables area, number of lanes, and speed limit, different classes of these variables were combined together to form distinct dummy variables. These dummy variables are defined in a way that each variable represents an area, number of lanes, and a speed limit (for example, "urban_3p_65" represent the segments located in urban areas with more than 3 lanes and a speed limit of 65 mph). Three categories were defined for the nolanes i.e., 2, 3, and 3+ in each direction, and four speed limit categories were chosen for speed limits of 55, 60, 65, and 70 mph. Overall, 12 categories for urban and 6 categories for rural areas were defined. Of course, the number of observations was not sufficient for the segments in rural areas with three lanes in each direction. Data for these rural categories, however, were kept in the analysis and considered with the rural_2_70 group as the base category. Prior to this action, their effect was tested statistically and found to be not significant. Therefore, the interpretation of the effect of the dummy variables with regards to the base category would not be affected. Additional dummy variables were defined for each season by interacting the dummy variables defined above with seasonal dummy variables. The new variables combined represent a season, an area type, number of lanes, and a speed limit (e.g. "winter_urban_3p_65"). Similar categories of number of lanes and speed limit were used to create dummy variables with seasonal interaction.

Further, the interaction of the main factors of the model with the area type and seasons were considered to examine their effects on crash frequency across area types and seasons. A dummy variable, "area" was set to 0 for rural and 1 for urban, and the interactions of this variable was defined with the main variables LnAADT, percent of heavy vehicles, lane width, and PSR. The variable "area" was confounded with the

www.manaraa.com

combinatory dummy variables such as “urban_3p_65” and, therefore, was not used in the model. Table 2 shows the interaction variables used in this study.

A continuous variable named “Safety_plan” was also defined to account for the MSHSP implementation through the years, 2005-2011. This variable takes the value of 0 for all the months of 2002 to 2004 and gradually increases from 0 to 1 for each month over the implementation period starting January 2005. The increase in the value of this variable coincides approximately with the proportion of the safety features completed at a given time (e.g. see MoDOT (2004) and MoDOT (2008) for more information on the objectives of the MSHSP).

A preliminary frequency analysis of the data was conducted in order to ensure that a sufficient number of observations (60 observations were considered satisfactory) were available. This was conducted to estimate the effect of each level of variables within each season using the analysis of repeated measurements by the GEE method (Mancl and DeRouen, 2001; Hutchings et al., 2003). Note that the dependent variable used in this study is the monthly crash count, therefore the above frequency analysis was carried out on a monthly basis.

www.manaraa.com

Table 1. Definition of the continuous and dummy variables considered for analysis

Variable[1]	Definition
Continuous variables	
LnAADT	Natural logarithm of annual average daily traffic in vehicles per day
PSR	Index representing pavement serviceability rate
Percentcommercial	Annual average percentage of trucks or heavy vehicles.
Congestionindex	Index representing congestion level
Lanewidth	Width of the highway lane in feet
Safety_plan	Proportion of the safety strategies implemented at a given time
Combinatory dummy variables[2]	
Urban_2_55	1, if area = Urban, number of lanes = 2, and speed limit = 55 mph, 0 otherwise
Urban_2_60	1, if area = Urban, number of lanes = 2, and speed limit = 60 mph, 0 otherwise
Urban_2_65	1, if area = Urban, number of lanes = 2, and speed limit = 65 mph, 0 otherwise
Urban_2_70	1, if area = Urban, number of lanes = 2, and speed limit = 70 mph, 0 otherwise
Urban_3_55	1, if area = Urban, number of lanes = 3, and speed limit = 55 mph, 0 otherwise
Urban_3_60	1, if area = Urban, number of lanes = 3, and speed limit = 60 mph, 0 otherwise
Urban_3_65	1, if area = Urban, number of lanes = 3, and speed limit = 65 mph, 0 otherwise
Urban_3_70	1, if area = Urban, number of lanes = 3, and speed limit = 70 mph, 0 otherwise
Urban_3p_55[3]	1, if area = Urban, number of lanes > 3, and speed limit = 55 mph, 0 otherwise
Urban_3p_60	1, if area = Urban, number of lanes > 3, and speed limit = 60 mph, 0 otherwise
Urban_3p_65	1, if area = Urban, number of lanes > 3, and speed limit = 65 mph, 0 otherwise
Urban_3p_70	1, if area = Urban, number of lanes > 3, and speed limit = 70 mph, 0 otherwise
Rural_2_60	1, if area = Rural, number of lanes = 2, and speed limit = 60 mph, 0 otherwise
Rural_2_65	1, if area = Rural, number of lanes = 2, and speed limit = 65 mph, 0 otherwise
Rural_2_70[5]	1, if area = Rural, number of lanes = 2, and speed limit = 70 mph, 0 otherwise
Seasonal dummy variables	
Spring_dummy	Indicator variable for spring season (1, if season[4] = "spring", 0 otherwise)
Summer_dummy	Indicator variable for summer season(1, if season = "summer", 0 otherwise)
Winter_dummy	Indicator variable for winter season (1, if season = "winter", 0 otherwise)

1. For a list of interaction variables see Table 2

2. These variables were defined in this format due to confounding effects of the incorporating variables. Variables in bold were used in the final model.

3. 3p means 3-plus indicating more than 3 lanes.

4. Spring season defined "March to May", summer as "June to August", fall as "September to November", and winter as "December to February"

5. This variable is considered as the base category for the other combinatory dummy variables and not directly used in the model; however, the interaction of this variable with the seasonal dummies is used in the model (see Table 2)

www.manaraa.com

Table 2. Definition of the interaction variables considered for analysis

Variable[1]	Definition (interaction between variables)
Area type interaction variables	
Areadt	"Area"[2] and "LnAADT"
Areapsr	"Area" and "PSR"
Areacommercial	"Area" and "Percentcommercial"
Areacongestion	"Area" and "Congestionindex"
Arealanewidth	"Area" and "Lanewidth"
AreaSafety_plan	"Area" and "Safety_plan"
Seasonal interaction variables[3]	
Season_area	seasonal dummy variable[3] and "Area" [2]
Season _lnAADT	seasonal dummy and "LnAADT"
Season _PSR	seasonal dummy variable and "PSR"
Season _percentcommercial	seasonal dummy variable and "Percentcommercial"
Season _congestion	seasonal dummy variable and "Congestionindex"
Season _lanewidth	seasonal dummy variable and "Lanewidth"
Season _Safety_plan	seasonal dummy variable and "Safety_plan"
Season _rural_2_60	seasonal dummy variable and the dummy "Rural_2_60"[1]
Season_rural_2_65	seasonal dummy variable and the dummy "Rural_2_65"
Season_rural_2_70	seasonal dummy variable and the dummy "Rural_2_70"
Season_urban_2_55	seasonal dummy variable and the dummy "Urban_2_55"
Season_urban_2_60	seasonal dummy variable and the dummy "Urban_2_60"
Season_urban_2_65	seasonal dummy variable and the dummy "Urban_2_65"
Season_urban_2_70	seasonal dummy variable and the dummy "Urban_2_70"
Season_urban_3_55	seasonal dummy variable and the dummy "Urban_3_55"
Season_urban_3_60	seasonal dummy variable and the dummy "Urban_3_60"
Season_urban_3_65	seasonal dummy variable and the dummy "Urban_3_65"
Season_urban_3_70	seasonal dummy variable and the dummy "Urban_3_70"
Season_urban_3p_55	seasonal dummy variable and the dummy "Urban_3p_55"
Season_urban_3p_60	seasonal dummy variable and the dummy "Urban_3p_60"
Season_urban_3p_65	seasonal dummy variable and the dummy "Urban_3p_65"

1. For a list of continuous or dummy variables see Table 1

2. "Area"=0 for rural, and =1 for urban areas; though, it was not used in the model due to confounding effect with combinatory variables (see Table 1)

3. Seasonal interaction terms were defined for each one of the seasonal dummy variables (see Table 1) but for the sake of brevity, only in this table, the term "Season" is substituted for all the season names of "spring", "summer", and "winter"

www.manaraa.com

4. RESULTS AND DISCUSSION

A negative binomial model was developed using the GEE method incorporating an autoregressive Type 1 correlation structure within each segment over ten years of data. Table 3 presents the results of the model estimates, standard errors of the coefficients, and variables that were found to be statistically significant in the final model. Variables that were not statistically significant in the model were removed from the analysis in a one-variable-at-a-time manner and the model was run again. The main factors that were not significant remained in the model if an associated interaction term was found to be significant. This helped correctly interpret the interaction parameters of the model (Nelder, 1977; Cox, 1984).

The positive coefficient estimate for LnAADT indicates that higher traffic volume relates to higher number of crashes, a trend commonly observed in the literature (Zhang et al., 2012; Roque and Cardoso, 2014). A negative estimate for the interaction of this variable with the area type (Areadt) indicates that the overall effect of LnAADT (over all seasons) is lower in urban areas compared to rural areas. A statistically significant seasonal interaction of a variable provides information on how the effect of that variable alters in a given season. The interaction terms of LnAADT with spring and winter dummy variables (Spring_lnAADT and Winter_lnAADT) were found to have a positive and negative estimate, respectively. This indicates that the impact of traffic volume in increasing crash frequency is higher in spring and lower in winter, when compared to the fall season. This points to cautious driving and lower speeds as traffic volume increases (Elvik et al., 2004; Aarts and van Schagen, 2006) during the winter compared to the warmer seasons such as spring and summer. This seasonal effect was not found for summer when compared to the fall season. It should be noted that the overall effects of such seasonal interaction of any season on crash frequency was considered together to determine the effect of the season on crash frequency, and presented at the end of the current section.

www.manaraa.com

Table 3. Negative binomial model parameter estimates

Parameter	Estimate	Standard Error	Pr > |Z|
Intercept	-35.45	1.4066	**<0.0001**
LnAADT	3.3607	0.1219	**<0.0001**
PSR	-0.009	0.0064	0.1582
Percentcommercial	-1.9996	0.3547	**<0.0001**
Congestionindex	0.1579	0.1335	0.2368
Safety_plan	-0.1854	0.0509	**0.0003**
Areadt	-2.566	0.1541	**<0.0001**
Areacommercial	1.9305	0.4522	**<0.0001**
Areacongestion	-0.3362	0.1402	**0.0165**
Urban_2_55	27.285	1.7470	**<0.0001**
Urban_2_60	26.4782	1.7293	**<0.0001**
Urban_2_65	26.8975	1.7241	**<0.0001**
Urban_2_70	26.5672	1.7146	**<0.0001**
Urban_3_55	27.3625	1.7503	**<0.0001**
Urban_3_60	26.7808	1.7535	**<0.0001**
Urban_3_65	26.6031	1.7484	**<0.0001**
Urban_3_70	25.9504	1.7381	**<0.0001**
Urban_3p_55	27.2925	1.7698	**<0.0001**
Urban_3p_60	26.9873	1.7661	**<0.0001**
Urban_3p_65	26.5842	1.7702	**<0.0001**
Urban_3p_70	26.1134	1.7668	**<0.0001**
Rural_2_60	0.6788	0.1727	**<0.0001**
Rural_2_65	1.532	0.3626	**<0.0001**
Spring_dummy	-0.3435	0.2498	0.1690
Spring_lnAADT	0.0671	0.0181	**0.0002**
Spring_PSR	-0.0125	0.0046	**0.0068**
Spring_urban_2_70	0.0815	0.0314	**0.0095**
Summer_dummy	0.3949	0.1373	**0.0040**
Summer_percentcommer	0.4588	0.0941	**<0.0001**
Summer_PSR	-0.013	0.0042	**0.0019**
Summer_Safety_plan	-0.2151	0.0292	**<0.0001**
Summer_urban_3_55	-0.1088	0.0306	**0.0004**
Winter_dummy	3.6691	0.3648	**<0.0001**
Winter_lnAADT	-0.3375	0.0350	**<0.0001**
Winter_PSR	-0.0114	0.0047	**0.0158**
Winter_Safety_plan	0.251	0.0323	**<0.0001**
Winter_urban_2_55	0.2911	0.0599	**<0.0001**
Winter_urban_2_65	0.1799	0.0511	**0.0004**
Winter_urban_3_55	0.2339	0.0534	**<0.0001**
Winter_urban_3_60	0.3067	0.0567	**<0.0001**
Winter_urban_3_65	0.372	0.0542	**<0.0001**
Winter_urban_3_70	0.2279	0.0895	**0.0109**
Winter_urban_3p_55	0.3625	0.0591	**<0.0001**
Winter_urban_3p_60	0.4104	0.0535	**<0.0001**
Winter_urban_3p_65	0.4258	0.0886	**<0.0001**

Note 1: The p-values are used to determine the statistical significance of the variables. Bold estimates are significant at 95% level of confidence (p-value < 0.05)

Note 2: For variable definitions see Tables 1 and 2

www.manaraa.com

The variable PSR and its interaction with the area type were not found to be significant, but the interactions of PSR with the seasonal dummy variables (Spring_PSR, Summer_PSR, and Winter_PSR) were found to be statistically significant with negative estimates. This shows that although the effect of the pavement condition is not significant in affecting crash frequency during fall, a pavement in better condition reduces the likelihood of crashes by varying degrees over the spring, summer, and winter seasons, compared to the fall season. This effect was found to be highest for spring, followed by summer and winter seasons. Anastasopoulos and Mannering (2011) and Buddhavarapu et al. (2013) considered similar pavement characteristics in their analysis on an aggregated dataset over several years and found similar results.

The negative estimate for the coefficient of percentcommercial indicates that higher percentage of heavy vehicles is associated with lower crash frequency. This shows that drivers reduce their speed and are cautious while traveling close to heavy vehicles (Carson and Mannering, 2001; Lao et al., 2014). The positive estimate of the interaction of this variable with area (areacommercial) indicates that trucks have an increasing effect on crash frequency in urban areas compared to rural. Although Khorashadi et al. (2005) found that trucks result in higher severity (or fatal) crashes in rural areas compared to urban areas, this study considered all crash severities. And the difference in results may be due to the drivers being relatively less experienced with driving around heavy vehicles in urban areas and their interaction with these vehicles increased the likelihood of a crash occurrence. Also, percentcommercial showed a positive interaction effect with the summer variable (summer_percentcommercial) indicating a positive impact on crash frequency associated with truck percentage during the summer. This can be due to higher percent of recreational vehicles –RVs (considered as heavy vehicles) as well as freight movement during the summer. Further research is warranted in this regard.

Congestionindex was not found to be statistically significant, however, its interaction with area type (areacongestion) was found significant with a negative sign indicating fewer crash occurrences in congested urban areas compared to rural areas. A possible explanation is that drivers are used to congestion in urban areas and are prepared to prevent collisions, whereas in rural areas it is not that common to face congestion

www.manaraa.com

especially on Interstate highways. Therefore, relative to urban areas, congestion in rural areas is more likely to cause crashes.

The safety_plan variable, designated to capture the effects of implementation of strategies during the years 2005-2011 (MoDOT, 2004, 2008), was found to be statistically significant. The negative sign of its estimated coefficient indicates a reduction in the crash frequency during the implementation years through its completion and that the safety improvement strategies were effective. For a more detailed analysis, the interested reader is referred to Mojtaba A Mohammadi et al. (2014) on the safety effects of MSHSP during the first phase i.e., 2005-2008. The interaction terms of this variable with summer and winter dummy variables (summer_safety_plan and winter_safety_plan) were found to have a negative and positive estimate, respectively. This indicates that compared to the fall (or spring season), the effect of MSHSP in lowering crash frequency was higher for the summer, but none in the winter. This may be due to the severity-types of crashes (fatal, serious injury, etc.), as the main objective of MSHSP was to reduce the fatal and severe injury crashes (MoDOT, 2004, 2008) and such crashes were reduced as a result of its implementation; however, an increase in less severe crashes (e.g. run-off-road collisions due to snow/ice) occurred during the winter in spite of the safety strategies implemented. Further research is needed in this respect to model crash frequency for various levels of crash severity and collision types.

The combinatory dummy variables defined by the interaction of the variables area type, number of lanes, and speed limit were all found to be significant with a positive estimate for all the terms related to urban areas and two of the terms related to rural areas (rural_2_60, and rural_2_65). The results for the rural areas indicate that the speed limit less than 70 mph has an increasing effect on crash frequency when the number of lanes equals two. It should be noted that rural_2_70 was used as the base condition for all the combinatory variables, and the results have to be interpreted with comparison to the base variable. For the urban areas, keeping two factors fixed and observing the change in the estimate for the third factor (e.g. fix area type and number of lanes, and change speed limit) presents how each contributing factor (within the combination) affects the crash frequency. Figure 1 shows the results of this investigation. A consistent trend was not found for the effect of number of lanes on crash frequency. It should be noted that

www.manaraa.com

Noland and Oh (2004) and L.-Y. Chang (2005) found that higher number of lanes result in lower crash frequency. Increasing speed limit results in lower crash frequency except when the number of lanes is two, which is a somewhat counterintuitive result. However, when the overall Interstate highway system is considered, it is known that the interstate highway system is the safest system in the United States compared to the US and State highway systems. And most of the rural interstate highways have two lanes in each direction.

From Table 3 only the relative significance of the effect of combinatory dummy variables on crash frequency compared to the base factor (i.e. rural_2_70) can be determined by the corresponding p-values for each variable. A variable changing from one level to another has an effect on crash frequency, which is determined by subtracting the estimates of the two levels. The statistical significance of this difference was tested separately. For example, to test the significance of the effect of changing the speed limit from 50 to 60 mph for a segment in an urban area with 2 lanes (changing urban_2_50 to urban_2_60). The variable urban_2_50 was redefined to take the value of 1 when nolanes=2, and SL=50 or SL=60. This variable would be applicable if there was no difference as a result of this change in speed limit, which is the null hypothesis of our test. If the model used this new variable together with urban_2_60, a low p-value (e.g. less than 0.05) for the effect of urban_2_60 will indicate that the null hypothesis is rejected (at 95% level of confidence) and there is a statistically significant effect as a result of changing the speed limit from 50 mph to 60 mph (for a segment in an urban area with 2 lanes in each direction). Table 4 shows the effect and statistical significance of all the possible combinations for the two variables, number of lanes and speed limit.

www.manaraa.com

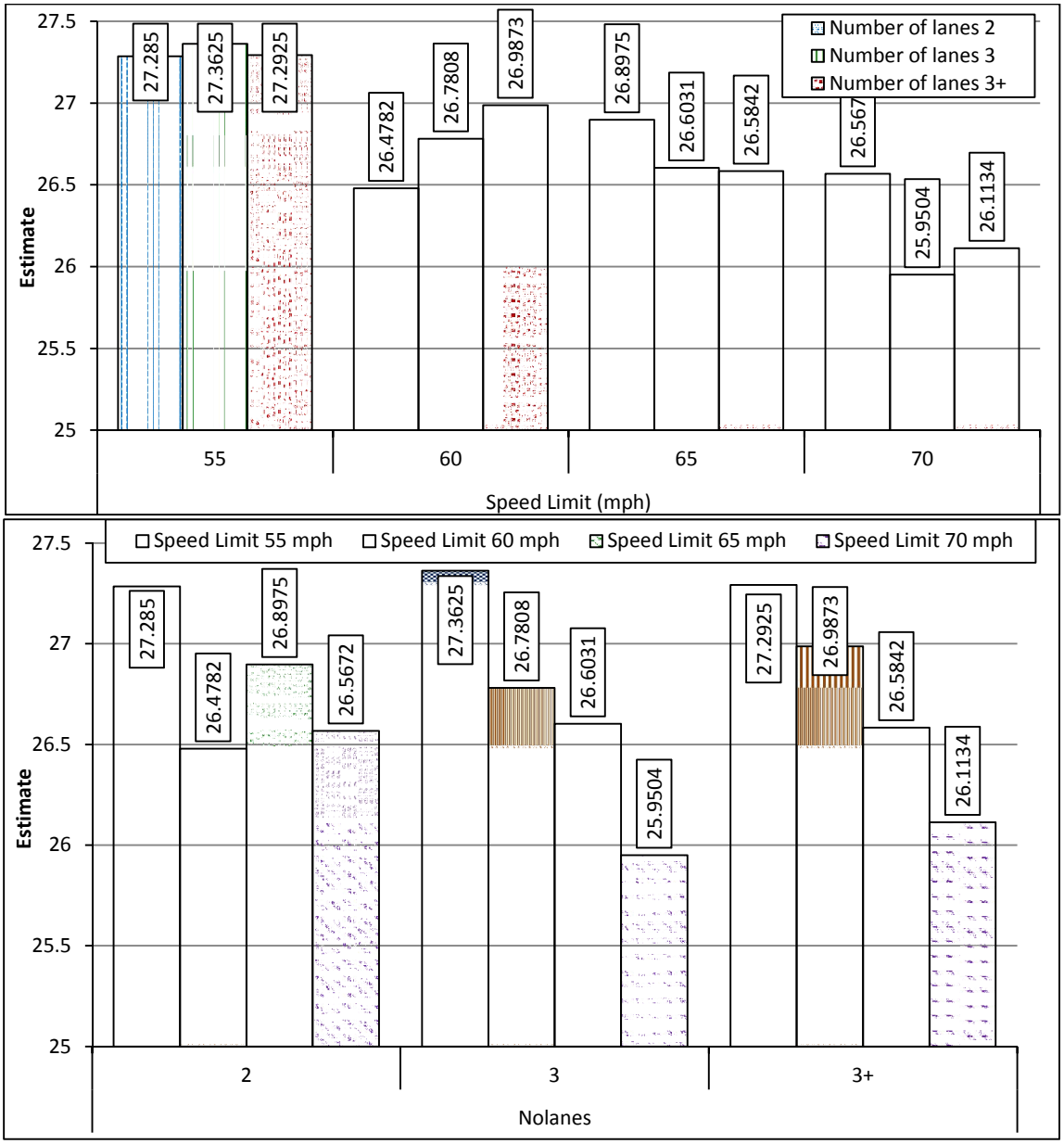

Figure 1. Estimates of the interaction terms between number of lanes, and speed limit in urban areas (top: effect of number of lanes within each class of speed limit, bottom: effect of speed limit within each class of number of lanes).

www.manaraa.com

Table 4. Overall amount and statistical significance of the effect of change in the number of lanes and speed limit on crash frequency in urban areas of Missouri

Nolanes		2				3				3+			
	SL	55	60	65	70	55	60	65	70	55	60	65	70
2	55	Est.	**-0.8069**	**-0.3875**	**-0.7178**	0.0775	**-0.5043**	**-0.682**	**-1.3347**	0.0075	**-0.2978**	**-0.7008**	**-1.1716**
		SD	0.1309	0.1151	0.1305	0.0942	0.0937	0.0948	0.1651	0.1225	0.0979	0.1504	0.2566
	60	0.8069	Est.	**0.4194**	0.089	**0.8844**	**0.3026**	0.1249	**-0.5278**	**0.8143**	**0.5091**	0.1061	-0.3647
		0.1309	SD	0.1073	0.0955	0.1066	0.1112	0.103	0.1628	0.1399	0.1163	0.1586	0.2573
	65			Est.	**-0.3304**	**0.465**	-0.1168	**-0.2945**	**-0.9472**	**0.395**	0.0897	**-0.3133**	**-0.7841**
				SD	0.09	0.0915	0.0955	0.0878	0.1527	0.1294	0.1035	0.1516	0.2526
	70				Est.	**0.7954**	0.2136	0.0359	**-0.6168**	**0.7253**	**0.4201**	0.0171	-0.4537
					SD	0.1069	0.1147	0.0996	0.1524	0.1449	0.1188	0.1583	0.254
3	55					Est.	**-0.5818**	**-0.7595**	**-1.4122**	-0.07	**-0.3753**	**-0.7783**	**-1.2491**
						SD	0.0606	0.0602	0.145	0.0967	0.064	0.1299	0.2453
	60						Est.	**-0.1777**	**-0.8304**	**0.5118**	**0.2065**	-0.1965	**-0.6673**
							SD	0.0608	0.1523	0.095	0.0629	0.1301	0.2465
	65							Est.	**-0.6527**	**0.6895**	**0.3842**	-0.0188	**-0.4896**
								SD	0.1446	0.0976	0.0621	0.1272	0.2439
	70								Est.	**1.3421**	**1.0369**	**0.6339**	0.1631
									SD	0.1698	0.1486	0.1851	0.2744
3+	55									Est.	**-0.3052**	**-0.7083**	**-1.1791**
										SD	0.0906	0.1458	0.2573
	60										Est.	**-0.403**	**-0.8738**
											SD	0.1211	0.2438
	65											Est.	-0.4708
												SD	0.2675
	70											0.4708	Est.
												0.2675	SD

Note. Nolanes, and SL represent the number of lanes and speed limit, respectively. *Est.* presents the model estimates (effect of change in nolanes or SL), and *SD* stands for estimate's standard error. Bold values are statistically significant at 95% level of confidence.

The two combinatory variables rural_2_60 and rural_2_65 were found to be statistically significant with positive coefficient estimates. This indicates that with two lanes in a rural area, the likelihood of crash occurrence becomes higher when the speed limit changes from 70 mph to 60 or 65 mph, with the latter change showing a higher effect (see coefficient estimates in Table 3). The effect of changing speed limit from 60 to 65 mph was tested separately, similar to the tests conducted for the urban area and this effect was not statistically significant. The highway sections with a speed of 60 or 65 mph were those with the highest crash frequency or hotspots and to make these vulnerable sections safer, Missouri DOT lowered the posted speed limits to 65 mph and at times to 60 mph (from 70 mph).

www.manaraa.com

The positive estimate of spring_urban_2_70 indicates that in the spring season reducing the number of lanes from more than three to two lanes and keeping the speed limit at 70 mph in urban areas will result in a higher crash frequency. The negative estimate for the Summer_urban_3_55 indicates that in the summer season a reduction in speed limit from 70 to 55 mph and number of lanes from more than three to three lanes will result in a significantly lower crash frequency in urban areas. Similar to the effect of change in speed limit and the number of lanes for urban areas, from Table 3 many of the interaction terms of these combinatory variables (urban_nolanes_sl) with the winter_dummy were found to be statistically significant. For each season, the combination of the highest level of the number of lanes (more than three) and highest level of speed limit (70 mph) was used as the base category. That is, for example, all the interaction terms found to be significant in the winter are compared with the base category for winter, which is winter_urban_3p_70. Figure 2 shows the result of coefficient estimates for these terms during the winter. Further, Table 5 presents the effect of the change in speed limit and/or number of lanes during the winter. A consistent trend was not observed as a result of these estimates except that a change in the number of lanes from three to more than three lanes while keeping the speed limit constant results in an increase in crash frequency during the winter season. Also, increasing speed limit to 65 mph showed an increase in crash frequency only when there were three or more than three lanes. These findings represent winter season in urban areas only.

www.manaraa.com

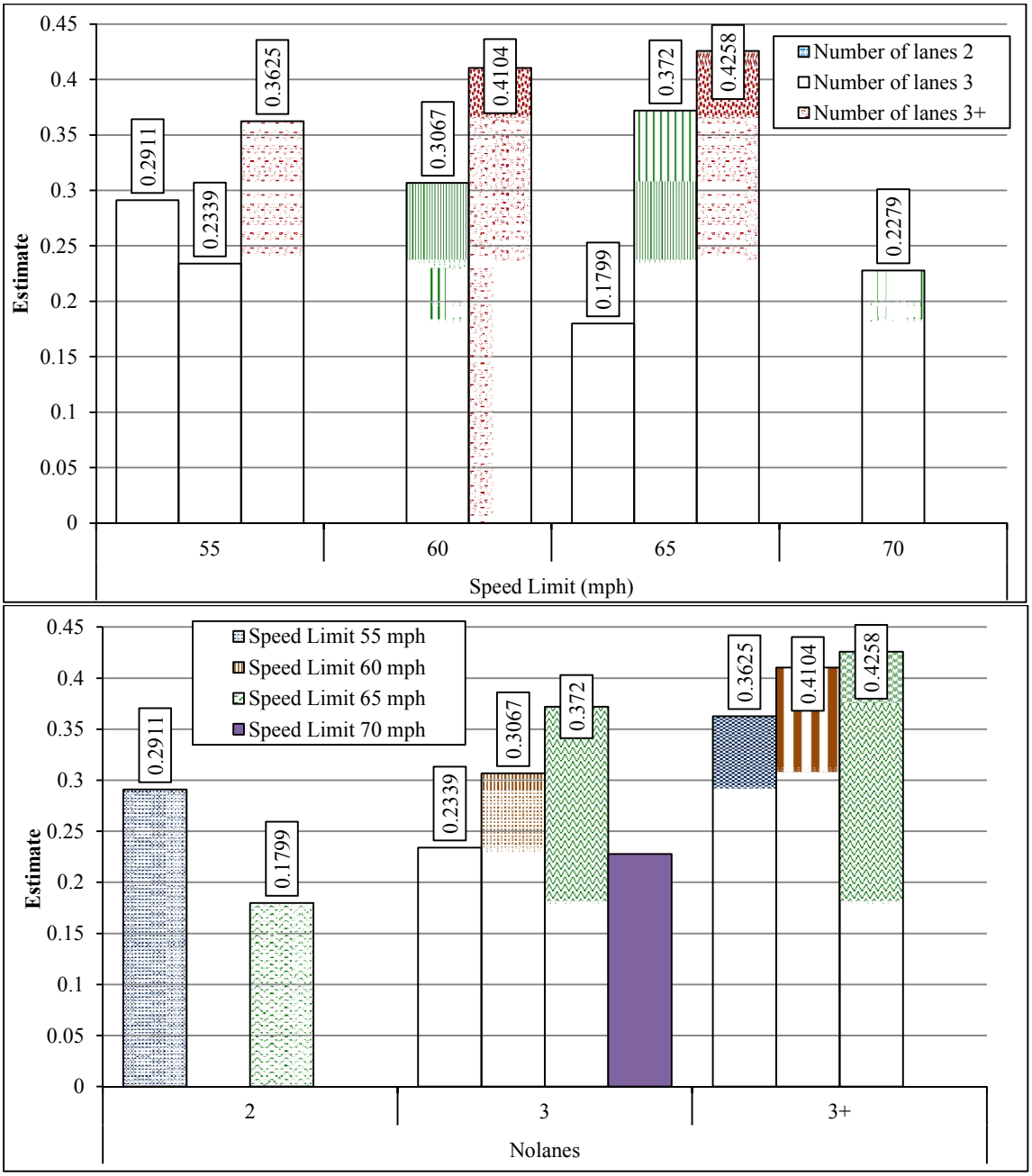

Figure 2. Estimates of the significant interaction terms between number of lanes, and speed limit in urban areas during winter season (top: effect of number of lanes within each class of speed limit, bottom: effect of speed limit within each class of number of lanes).

www.manaraa.com

Table 5. Amount and statistical significance of the effect of change in the number of lanes and speed limit on crash frequency in urban areas during the winter season

Nolanes		2		3			3+			
	SL	55	65	55	60	65	70	55	60	65
2	55	Est	-0.1112	-0.0573	**0.1525**	0.0809	0.1459	0.0714	**0.2577**	0.1346
		SD	0.0642	0.0532	0.0533	0.0579	0.0909	0.0559	0.0503	0.0893
	65		Est	0.0539	**0.2218**	**0.1921**	0.1768	**0.1826**	**0.326**	**0.2458**
			SD	0.0581	0.056	0.0608	0.0908	0.0615	0.0528	0.0917
3	55			Est	0.0728	**0.1382**	-0.0059	**0.1287**	**0.1765**	**0.1919**
				SD	0.0449	0.0495	0.0928	0.047	0.0408	0.0843
	60				Est	0.0653	-0.0788	0.0558	**0.1037**	0.1191
					SD	0.0516	0.0948	0.0472	0.0404	0.0846
	65					Est	-0.1441	-0.0095	0.0384	0.0538
						SD	0.0951	0.0526	0.0471	0.0872

From the estimates of the seasonal variables summer_dummy and winter_dummy, it can be stated that the summer and winter seasons affect crash frequency differently than the fall season and this effect is significant. The coefficient estimates of these variables, however, must be interpreted together with the estimates for their interactions with other main factors of the model. Since there are also negative coefficient estimates in the model for statistically significant interaction terms related to the summer and winter seasons, the positive sign of the estimate for these seasons does not necessarily mean that there are higher crash occurrences in those seasons. Similarly, statistically significant interaction terms related to the spring season were found in the model that affects the interpretation of the negative sign found for the estimated coefficient for spring_dummy variable, however, it was not found to be a significant factor in the model.

In order to find the effect of a season, only those variables related to the season of interest were considered in the model to predict the crash frequency. For example, to determine the effect of summer season in crash frequency the variables summer_dummy, summer_percentcommerial, summer_PSR, summer_safety_plan, and summer_urban_3_55 were used to predict the number of crashes for each segment. The sum of these predicted values is the crash frequency over the period of analysis attributable only to the specific seasonal effect (summer). This value provides a criterion for comparison of the overall effect of each season. The sum of the predicted crash

www.manaraa.com

frequency values for all seasonal effects were calculated and compared. All of the crash predictions were found to be positive values with the highest value for winter followed by summer and spring. Note that these effects are statistically significant as they were calculated using those significant factors of the model.

5. CONCLUSIONS AND RECOMMENDATIONS

The objective of this study was to investigate the seasonal effects on crash causality factors. A longitudinal negative binomial model was developed using the generalized estimating equation (GEE) method with autoregressive Type 1 correlation structure. The study used crash data from 2002-2011 for six main Interstate highways in Missouri. This study also evaluated the effects of Missouri's Strategic Highway Safety Plan (MSHSP) on crash frequency using a relatively simple approach.

The natural logarithm of AADT (LnAADT) was found to be statistically significant with a positive estimate indicating higher traffic volume results in higher number of crashes. The negative estimate for the interaction term, LnAADT and area type showed that the traffic volume had a smaller effect on urban areas compared to rural areas. Furthermore, LnAADT had statistically significant interaction terms with spring and winter seasons with positive and negative estimates, respectively. This indicates that, compared to the fall season, traffic volume had a higher effect in increasing the crash occurrence in spring and lower in winter time. Such seasonal effect of traffic volume was not found to be significant for the summer compared to the fall season.

Pavement serviceability rating (PSR) was not found to be statistically significant, but had significant interactions with the seasonal variables in the model. This indicated that a better quality pavement reduces the likelihood of crash occurrence by varying degrees over the seasons of spring, summer, and winter, compared to the fall season. This crash reducing effect was highest for spring season followed by summer and winter but the difference between the effects is minor.

Similar to PSR, only interaction terms of Congestion Index with area type was significant indicating a decrease in crash frequency associated with congestion in urban areas compared to rural areas. This indicated that drivers do not commonly experience

www.manaraa.com

congestion in rural areas especially on Interstate highways, hence, it is more likely to cause crashes relative to urban areas.

Higher percentage of heavy vehicles showed reduction in crash frequency. This effect was found to be higher in urban areas. Also, this variable showed a positive interaction effect with the summer variable indicating higher crash frequency associated with truck percentage during the summer. This indicated increase in travel during the warmer season and as a result, higher frequency of crashes. This result is similar to the LnAADT variable; as traffic increases the frequency of crashes also increase.

In terms of the effects of number of lanes and speed limit on crash frequency, this study did not show a consistent trend. This result is contrary to the findings of Noland and Oh (2004) and L.-Y. Chang (2005), who found that higher number of lanes results in lower crash frequency. A more in-depth analysis is therefore required to explain this behavior.

A significant difference was found in the effect of winter and summer seasons on crash frequency compared to fall season. The model estimates for the seasonal variables show that the summer and winter seasons affect crash frequency significantly different than the fall season. Spring season was not found to be a significant factor in the model. Considering all the main and interaction seasonal terms in the model and analyzing for the effects of seasons on crash frequency, the results indicate that summer, spring and winter seasons have an increasing effect on the crash frequency compared to the fall season. Winter season had the highest effect in positively affecting crash occurrences followed by summer and spring. Many of the interaction terms defined by the area type, number of lanes, and speed limit were found to be statistically significant only in the winter season but a consistent trend in their estimated values was not observed.

A significant difference was found in the effects of winter and summer seasons on crash frequency compared to the fall season. The model estimates for the seasonal variables show that the summer and winter seasons affect crash frequency significantly different than the fall season. Considering all of the main and interaction terms in the model and analyzing for the effects of seasons on crash frequency, the results indicate that summer, spring and winter seasons have an increasing effect on the crash frequency compared to the fall season. Winter season had the highest effect in positively affecting

www.manaraa.com

crash occurrences followed by summer and spring. Many of the interaction terms defined by the area type, number of lanes, and speed limit were found to be statistically significant only in the winter season, but a consistent trend in their estimated values was not observed.

Safety_plan defined the effectiveness of the MSHSP and was found to be statistically significant with a negative estimate. This indicates that MSHSP effectively reduced the crash frequency during the years of implementation and similar strategic plans should be promoted as an effective way to reduce crashes. Other studies also support the findings of this study (Jung et al., 2013b). Safety_plan showed statistically significant negative and positive interaction with the summer and winter variables, respectively. This indicated that MSHSP had a larger effect in reducing the crashes during the summer and smaller effect during the winter, compared to the fall (or spring seasons). One of the objectives of the MSHSP was to reduce fatal and severe injury crashes (MoDOT, 2004, 2008). It is possible that these safety improvement strategies reduced severe crashes during the winter, but with an increase in less severe crashes. Crash frequency specifically in terms of crash severity requires further research. Also, Mojtaba A Mohammadi et al. (2014) conducted a more detailed study on the effects of MSHSP (during the first phase, i.e. 2005-2008) on crash frequency by various collision types and severity. The availability of more detailed data for specific countermeasures (such as adding median barriers, rumble strips, etc.) will certainly provide detailed understanding of how certain measures affect crash statistics.

The developed model in this paper enhances the understanding of seasonal crash patterns and whether the magnitude and/or types of various effects are different according to climatic changes. The results of this study will help in better identification of crash countermeasures with regards to the different times of the year.

7. REFERENCES

1. Elvik, R. How much do road accidents cost the national economy? *Accident Analysis & Prevention,* Vol. 32, No. 6, 2000, pp. 849-851.
2. NHTSA. *Traffic safety facts*, 2008.

www.manaraa.com

3. NHTSA. *Traffic safety facts*, 2009.
4. Peden, M., R. Scurfield, D. Sleet, D. Mohan, A. A. Hyder, E. Jarawan, and C. D. Mathers. World report on road traffic injury prevention. World Health Organization Geneva, 2004.
5. HSM. *Highway Safety Manual*. AASHTO, 2010.
6. Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice,* Vol. 44, No. 5, 2010, pp. 291-305.
7. Garber, N., and L. Hoel. *Traffic & highway engineering*. Cengage Learning, 2008.
8. Hilton, B. N., T. A. Horan, R. Burkhard, and B. Schooley. SafeRoadMaps: Communication of location and density of traffic fatalities through spatial visualization and heat map analysis. *Information Visualization,* Vol. 10, No. 1, 2011, pp. 82-96.
9. Ahmed, M., H. Huang, M. Abdel-Aty, and B. Guevara. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident analysis and prevention,* Vol. 43, No. 4, 2011, pp. 1581-1589.
10. Yu, R., M. Abdel-Aty, and M. Ahmed. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident analysis and prevention,* Vol. 50, 2013, pp. 371-376.
11. Yang, H., K. Ozbay, O. Ozturk, and M. Yildirimoglu. Modeling work zone crash frequency by quantifying measurement errors in work zone length. *Accident analysis and prevention,* Vol. 55, 2013, pp. 192-201.
12. CERS. *National Rural Road Safety Public Opinion Survey*. http://www.ruralsafety.umn.edu/publications/nationalsafetysurvey/index.html. Accessed July 24, 2014.
13. Carson, J., and F. Mannering. The effect of ice warning signs on ice-accident frequencies and severities. *Accident analysis and prevention,* Vol. 33, No. 1, 2001, pp. 99-109.

www.manaraa.com

14. Zeger, S. L., and K.-Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 1986, pp. 121-130.
15. Wang, X., and M. Abdel-Aty. Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accident analysis and prevention,* Vol. 38, No. 6, 2006, pp. 1137-1150.
16. Lord, D., and B. N. Persaud. Accident prediction models with and without trend: application of the generalized estimating equations procedure. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1717*, 2000. pp. 102-108.
17. Giuffrè, O., A. Granà, T. Giuffrè, and R. Marino. Improving reliability of road safety estimates based on high correlated accident counts. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2019*, 2007. pp. 197-204.
18. Zorn, C. J. Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 2001, pp. 470-490.
19. Mohammadi, M. A., V. A. Samaranayake, and G. Bham. Crash Frequency Modeling using Negative Binomial Models: An Application of Generalized Estimating Equation to Longitudinal Data. *Accepted for publication in Analytic Methods in Accident Research,* Vol. 2, 2014.
20. MoDOT. *Missouri's blueprint for safer roadways.* Missouri Coalition for Roadway Safety. http://www.ite.org/safety/stateprograms/Missouri_SHSP.pdf. Accessed July 26, 2014.
21. MoDOT. *Missouri's blueprint to arrive alive.* Missouri Coalition for Roadway Safety. http://www.savemolives.com/documents/FINALBlueprintdocument.pdf. Accessed July 26, 2014.
22. Hutchings, C. B., S. Knight, and J. C. Reading. The use of generalized estimating equations in the analysis of motor vehicle crash data. *Accident analysis and prevention,* Vol. 35, No. 1, 2003, pp. 3-8.
23. Mancl, L. A., and T. A. DeRouen. A covariance estimator for GEE with improved small-sample properties. *Biometrics,* Vol. 57, No. 1, 2001, pp. 126-134.

www.manaraa.com

24. Nelder, J. A. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, 1977, pp. 48-77.
25. Cox, D. R. Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1984, pp. 1-24.
26. Zhang, Y., Y. Xie, and L. Li. Crash frequency analysis of different types of urban roadway segments using generalized additive model. *Journal of Safety research,* Vol. 43, No. 2, 2012, pp. 107-114.
27. Roque, C., and J. L. Cardoso. Investigating the relationship between run-off-the-road crash frequency and traffic flow through different functional forms. *Accident Analysis & Prevention,* Vol. 63, 2014, pp. 121-132.
28. Aarts, L., and I. van Schagen. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention,* Vol. 38, No. 2, 2006, pp. 215-224.
29. Elvik, R., P. Christensen, and A. Amundsen. Speed and road accidents. *An evaluation of the Power Model. TØI report,* Vol. 740, 2004, p. 2004.
30. Anastasopoulos, P. C., and F. L. Mannering. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. *Accident Analysis & Prevention,* Vol. 43, No. 3, 2011, pp. 1140-1147.
31. Buddhavarapu, P., A. Banerjee, and J. A. Prozzi. Influence of pavement condition on horizontal curve safety. *Accident Analysis & Prevention,* Vol. 52, 2013, pp. 9-18.
32. Lao, Y., G. Zhang, Y. Wang, and J. Milton. Generalized nonlinear models for rear-end crash risk analysis. *Accident analysis and prevention,* Vol. 62, 2014, pp. 9-16.
33. Khorashadi, A., D. Niemeier, V. Shankar, and F. Mannering. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. *Accident Analysis & Prevention,* Vol. 37, No. 5, 2005, pp. 910-921.
34. Mohammadi, M. A., V. Samaranayake, and G. H. Bham. Safety Effect of Missouri's Strategic Highway Safety Plan-Missouri's Blueprint for Safer Roadways. In *Transportation Research Board 93rd Annual Meeting*, 2014.

www.manaraa.com

35. Noland, R. B., and L. Oh. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. *Accident analysis and prevention,* Vol. 36, No. 4, 2004, pp. 525-532.
36. Chang, L.-Y. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science,* Vol. 43, No. 8, 2005, pp. 541-557.
37. Mohammadi, M. A. Longitudinal analysis of crash frequency data., Missouri University of Sicence and Technology, 2014.
38. Jung, S., Q. Xiao, and Y. Yoon. Evaluation of motorcycle safety strategies using the severity of injuries. *Accident analysis and prevention,* Vol. 59, 2013, pp. 357-364.

www.manaraa.com

SECTION

2. CONCLUSIONS

The objectives of this study were developed and presented in three different steps and the results of the analyses conducted in steps one to three were presented in the sections "Paper I", "Paper II", and "Paper III".

The first objective of this study was to develop a traditional negative binomial (NB) regression model using the maximum likelihood technique that overcomes the limitations of the regression to the mean phenomenon (Barnett et al., 2005) in before/after crash evaluation studies and does not have the complexity of the empirical Bayesian models which, comparatively, requires training and practice in addition to extensive data (Hauer, 1997; Persaud et al., 2004; Guo et al., 2010b; Shively et al., 2010; Yu et al., 2013b). A simple approach was introduced to address the issues mentioned above in evaluation of the effectiveness for the Missouri Strategic Highway Safety Plan (MSHSP) over the years 2005-2007. Several negative binomial regression models were developed for crash frequency of various collision types and severity levels. These models considered the frequency observations for the before-through-change conditions and accounted for the safety plan progression using a continuous variable that was set to zero for pre-implementation years and gradually increased to one over the implementation years. The results of the first part of this study (presented in the section "Paper I") show that the MSHSP was a successful policy by reaching its primary goal, that is, to reduce the frequency and severity of serious injury crash types. This study found a significant reduction of 10% in all crashes combined. These strategies had the highest effect on the fatal crashes (30% reduction) and particularly on the head-on crashes (37% reduction) that result the most fatalities. The results were also found to be consistent with other studies and suggest that the safety strategic plans should be promoted (Kempton et al., 2006; Jung et al., 2013a).

The next part of this study (presented in the section "Paper II") uses more years of data to increase the reliability of the frequency models developed. More years of data reduces the standard errors of the model estimates but creates a serial correlation in

www.manaraa.com

repeated observations of a highway segment over the years (Park and Lord, 2009; Castro et al., 2012; Bhat et al., 2014; Mannering and Bhat, 2014; Zou et al., 2014). This correlation results in the underestimation of standard errors and subsequently biased model estimates (Ulfarsson and Shankar, 2003; Washington et al., 2011; Dupont et al., 2013; Mohammadi et al., 2013; Bhat et al., 2014; Xiong et al., 2014). The second part of this study aims to address this problem of biased model estimates due to temporal correlation in repeated observations. A longitudinal negative binomial model was developed using generalized estimating equations (GEE) technique to model ten years (2002-2011) of crash frequency data in Missouri. The GEE method used in this study accounts for the temporal correlation of repeatedly measured frequency data.

An autoregressive Type 1 structure was determined to be an appropriate correlation structure for the data. The results of the model using GEE method were then compared with the results of a traditional NB model using the maximum likelihood estimation (MLE) method. It was found that the GEE model, allowing for temporal correlations is a superior model comparatively providing more accurate and less biased estimates which is agreement with the literature (Lord and Persaud, 2000; Ulfarsson and Shankar, 2003; Mannering and Bhat, 2014).

The natural logarithm of AADT (LnAADT) found to be a statistically significant factor in the model with a positive estimate, indicating higher number of crashes with higher traffic volume. Also, the negative estimate for the interaction term of LnAADT and area type showed that the traffic volume have a smaller effect in urban than rural areas. An increase in speed limit was found to result in a decrease in crash frequency in urban areas (a somewhat counterintuitive result which may be related to the reverse association of crash occurrence with the speed limit), and the change in the number of lanes did not show a consistent trend in affecting the crash frequency.

The results show that by considering temporal correlation in the model (using GEE technique), some variables may become insignificant. Percent commercial with a negative estimate was only found to be significant in the MLE model indicating that heavy vehicles result in fewer crashes. This indicates drivers may use more caution and reduce their speed when they travel close to large vehicles. Carson and Mannering (2001) and Lao et al. (2014) also found similar results of the effect of truck percentage on crash

www.manaraa.com

occurrences. Also, the positive sign of the significant interaction of this variable with the area type indicated that the impact of commercial vehicles on crash frequency is higher in urban areas. Another factor that was only found to be significant in the traditionally estimated model (using MLE method) was lane width with a positive sign. This result, however, seems to be counterintuitive (see Li et al. (2008) and Manuel et al. (2014) whose results are inconsistent with this study); however, other recent studies have found similar results to this study regarding the effect of lane width (Aguero-Valverde and Jovanis, 2009; Dong, Clarke, Richards, et al., 2014; Dong, Clarke, Yan, et al., 2014). Martens et al. (1997) note that when the lane widths decrease drivers show improved lane-keeping and reduce their speed. These inconsistent conclusions indicate that further study of this matter may be required. Additionally, the autoregressive Type 1 correlation structure was found to be an appropriate structure for this type of data. The results of this analysis suggest that if crash data is available for several years, it is recommended to use larger data sets to increase the model reliability, but also to account for the temporal correlations in the data. This provides more accurate models and therefore, safety policies and crash countermeasures based on such models will be more efficient in saving lives and resources. This study confirms that the use of GEE is a good approach for addressing the serial correlation in crash frequency data.

The third part of this study seeks to further enhance the applicability of the crash prediction models by investigating the seasonal effects on crash causality factors. The few studies found in this regard in the literature (Carson and Mannering, 2001; Ahmed et al., 2011; Hilton et al., 2011; Yu et al., 2013a) suggest that policy-makers can improve the safety of specific roadway segments according to seasonal changes of the effect of crash factors. The last part of this study presents an in-depth analysis of the seasonality of crash causes by developing a longitudinal negative binomial model using ten years of crash data on six main interstate highways of Missouri. This analysis also uses generalized estimating equation (GEE) technique to develop the model. The statistical significance of the interaction of the main crash factors with the seasonal variables were examined in the model. The effects of interventions made by the Missouri Strategic Highway Safety Plan (MSHSP) over the years 2005-2011 is also investigated. The results

www.manaraa.com

(presented in the section "Paper III") provide a better understanding of the change in the effect of crash causes over different seasons in a year.

Similar to the results of the previous part, the natural logarithm of AADT (LnAADT) and its interaction with area type were found to be statistically significant with a positive and negative estimate, respectively. This indicates higher traffic volume results in higher number of crashes and this effect of traffic volume on crash occurrence is smaller in urban areas compared to rural. Furthermore, LnAADT had statistically significant interaction terms with spring and winter seasons with positive and negative estimates, respectively. This indicates that, compared to the fall season, traffic volume has a higher effect in increasing the crash occurrence in spring and lower in winter time. Results indicate that better quality pavement reduces the crash occurrence differently over the seasons; this crash reducing effect was highest for spring season followed by summer and winter. The significant interaction of congestion index with area type indicates a decrease in crash frequency associated with congestion in urban areas compared to rural areas. This suggests that drivers do not commonly expect congestion in rural areas especially on interstate highways, hence, it is more likely to cause crashes relative to urban areas. Higher truck percentage showed reduction in crash frequency and this effect was found to be higher in urban areas. Also, this variable showed a positive interaction effect with summer variable indicating higher crash frequency associated with truck percentage during summer time. This may be due to the increase in recreational driving in the warmer season and as a result, higher frequency of crashes. No consistent trend was found in terms of the effects of number of lanes and speed limit on crash frequency. This result is contrary to the findings of Noland and Oh (2004) and L.-Y. Chang (2005), who found that higher number of lanes results in lower crash frequency. A more in-depth analysis is therefore required to explain this behavior. The results indicate that winter season had the highest effect in positively affecting crash occurrences followed by summer and spring. Many of the interaction terms defined by the area type, number of lanes, and speed limit were found to be statistically significant only in the winter season, but a consistent trend in their estimated values was not observed. Results show that the MSHSP effectively reduced the crash frequency during the years of implementation and similar strategic plans should be promoted as an effective way to

www.manaraa.com

reduce crashes. It was also found that MSHSP had a larger effect in reducing the crashes during the summer and smaller effect during the winter, compared to the fall (or spring seasons). One of the objectives of the MSHSP was to reduce fatal and severe injury crashes (MoDOT, 2004, 2008). It is possible that these safety improvement strategies reduced severe crashes during the winter, but with an increase in less severe crashes. Crash frequency specifically in terms of crash severity requires further research.

Further analysis of the effectiveness of particular SHSP that focus on the specific emphasis areas (such as adding median barriers, rumble strips, etc.) identified in the SHSP is warranted in future studies to obtain a more detailed understanding of how the implementation of specific safety measures affect safety. Provided the specific implementation data on the highways are available, future studies will consider examination of the effect of safety improvement plans (such as 'adding median barrier') on the type and injury severity of crashes.

www.manaraa.com

APPENDIX A.

MATLAB ALGORITHM FOR READING THE CRASH DATA BASE, ROAD INVENTORY DATA BASE, ASSIGNING SEGMENT IDENTIFICATIONS, AGGREGATING YEARLY, MONTHLY, AND SEASONAL CRASH FREQUENCY

www.manaraa.com

```
%% ############################################################
% ################ The State and US highways reading ##########
% ############################################################
% the file below cannot be read in one run, it was divided into three
% sheets so the MATLAB could read it and perform the tasks

data=xlsread('final MO-US 1999-2012 ready','mixed valids');

index_rural=data(:,15)==1;
data_rural=data(index_rural,:);

index_urban=or(data(:,15)==2, data(:,15)==3);
data_urban=data(index_urban,:);

index_rural_divided=and(data(:,15)==1, data(:,16)==0);
data_rural_divided=data(index_rural_divided,:);

index_rural_undivided=and(data(:,15)==1, data(:,16)==1);
data_rural_undivided=data(index_rural_undivided,:);

index_urban_divided=and(or(data(:,15)==2, data(:,15)==3),
data(:,16)==0);
data_urban_divided=data(index_urban_divided,:);

index_urban_undivided=and(or(data(:,15)==2, data(:,15)==3),
data(:,16)==1);
data_urban_undivided=data(index_urban_undivided,:);

SAVEPATH=strcat(pwd,filesep,'matlab');
if ( ~isdir(SAVEPATH))
    mkdir(SAVEPATH);
end
SAVEFILENAME=strcat(SAVEPATH,filesep,'mixed_valid_mous.mat');
disp(['The count data was saved to: ',SAVEFILENAME]);
save(SAVEFILENAME,
'data','data_rural','data_urban','data_rural_divided',...

'data_rural_undivided','data_urban_divided','data_urban_undivided');

clear index_rural index_urban SAVEFILENAME SAVEPATH
index_rural_divided...
        index_rural_undivided index_urban_divided
index_urban_undivided;


%% ############################################################
% ################ The Interstate highways reading ########
% ############################################################
% reading data from excel and saving into a .mat file along
% with the data divided for light and dark time

data=xlsread('final interstates 1999-2012 ready.xlsx','mixed valids');
indexlight0=data(:,46)==0;
```

www.manaraa.com

```
indexlight1=data(:,46)==1;
indexweather1=data(:,47)==1;
indexweather2=data(:,47)==2;
indexweather3=data(:,47)==3;
indexseverefatal=data(:,48)==1;
indexseverenotfatal=data(:,48)==0;
indexseverefataldisable=data(:,49)==1;
indexseverenotfataldisable=data(:,49)==0;

data_dark=data(indexlight0,:);
data_light=data(indexlight1,:);
data_cold=data(indexweather2,:);
data_rain=data(indexweather3,:);
data_clear=data(indexweather1,:);
data_severe_fatal=data(indexseverefatal,:);
data_severe_notfatal=data(indexseverenotfatal,:);
data_severe_fataldisable=data(indexseverefataldisable,:);
data_severe_notfataldisable=data(indexseverenotfataldisable,:);

indexlight0weather1severefatal= (data(:,46)==0 & data(:,47)==1 &
data(:,48)==1);
datalight0weather1severefatal= data(indexlight0weather1severefatal,:);

indexlight0weather1severenotfatal= (data(:,46)==0 & data(:,47)==1 &
data(:,48)==0);
datalight0weather1severenotfatal=
data(indexlight0weather1severenotfatal,:);

indexlight0weather2severefatal= (data(:,46)==0 & data(:,47)==2 &
data(:,48)==1);
datalight0weather2severefatal= data(indexlight0weather2severefatal,:);

indexlight0weather2severenotfatal= (data(:,46)==0 & data(:,47)==2 &
data(:,48)==0);
datalight0weather2severenotfatal=
data(indexlight0weather2severenotfatal,:);

indexlight0weather3severefatal= (data(:,46)==0 & data(:,47)==3 &
data(:,48)==1);
datalight0weather3severefatal= data(indexlight0weather3severefatal,:);

indexlight0weather3severenotfatal= (data(:,46)==0 & data(:,47)==3 &
data(:,48)==0);
datalight0weather3severenotfatal=
data(indexlight0weather3severenotfatal,:);


indexlight1weather1severefatal= (data(:,46)==1 & data(:,47)==1 &
data(:,48)==1);
datalight1weather1severefatal= data(indexlight1weather1severefatal,:);

indexlight1weather1severenotfatal= (data(:,46)==1 & data(:,47)==1 &
data(:,48)==0);
```

www.manaraa.com

```
datalight1weather1severenotfatal=
data(indexlight1weather1severenotfatal,:);

indexlight1weather2severefatal= (data(:,46)==1 & data(:,47)==2 &
data(:,48)==1);
datalight1weather2severefatal= data(indexlight1weather2severefatal,:);

indexlight1weather2severenotfatal= (data(:,46)==1 & data(:,47)==2 &
data(:,48)==0);
datalight1weather2severenotfatal=
data(indexlight1weather2severenotfatal,:);

indexlight1weather3severefatal= (data(:,46)==1 & data(:,47)==3 &
data(:,48)==1);
datalight1weather3severefatal= data(indexlight1weather3severefatal,:);

indexlight1weather3severenotfatal= (data(:,46)==1 & data(:,47)==3 &
data(:,48)==0);
datalight1weather3severenotfatal=
data(indexlight1weather3severenotfatal,:);


SAVEPATH=strcat(pwd,filesep,'matlab');
if ( ~isdir(SAVEPATH))
    mkdir(SAVEPATH);
end
SAVEFILENAME=strcat(SAVEPATH,filesep,'mixed_valid_interstates.mat');
disp(['The count data was saved to: ',SAVEFILENAME]);
save(SAVEFILENAME, 'data', 'data_dark', 'data_light', 'data_cold',
'data_rain', 'data_clear', ...
                    'data_severe_fatal', 'data_severe_notfatal',
'data_severe_fataldisable', ...
                    'data_severe_notfataldisable',
'datalight0weather1severefatal', ...
                    'datalight0weather1severenotfatal',
'datalight0weather2severefatal', ...
                    'datalight0weather2severenotfatal',
'datalight0weather3severefatal', ...
                    'datalight0weather3severenotfatal',
'datalight1weather1severefatal', ...
                    'datalight1weather1severenotfatal',
'datalight1weather2severefatal', ...
                    'datalight1weather2severenotfatal',
'datalight1weather3severefatal', ...
                    'datalight1weather3severenotfatal');

clear indexlight0 indexlight1 SAVEFILENAME SAVEPATH indexweather1
indexweather2 indexweather3 ...
    indexseverefatal indexseverenotfatal indexseverefataldisable
indexseverenotfataldisable ...
    indexlight0weather1severefatal indexlight0weather1severenotfatal
indexlight0weather2severefatal ...
    indexlight0weather2severenotfatal indexlight0weather3severefatal
indexlight0weather3severenotfatal...
```

www.manaraa.com

```
    indexlight1weather1severefatal indexlight1weather1severenotfatal
indexlight1weather2severefatal ...
    indexlight1weather2severenotfatal indexlight1weather3severefatal
indexlight1weather3severenotfatal;



%% ###############################################################
% ################ The Interstate Segment data reading #########
% ###############################################################
% reading the segment data and saving it into a .mat file
% to make it easy to access next time
% ---------------------------------------------------------------
SegmentData=xlsread('interstates segment data','mixed');

SAVEPATH=strcat(pwd,filesep,'matlab');
if ( ~isdir(SAVEPATH))
    mkdir(SAVEPATH);
end
SAVEFILENAME=strcat(SAVEPATH,filesep,'InterstateSegmentData.mat');
disp(['Segment data was saved into: ',SAVEFILENAME]);
save(SAVEFILENAME, 'SegmentData');


%% ###############################################################
% ################ Preparing data for analysis in SAS #########
% ###############################################################
clc; clear all; close all; format long; tic;
data=load('mixed_valid_interstates.mat');
segdata=load('segments_data_numbered.mat');
crashdata=data.data;
segdata=segdata.data3;

yearindex_crashdata=6;
monthindex_crashdata=7;
highwayname_index_crashdata=3;
travelwayid_index_crashdata=4;
novariables=size(data,2);
logindex_crashdata=12;
highwayname_index_segdata=1;
travelwayid_index_segdata=2;
yearindex_segdata=3;
beglogindex_segdata=6;
endlogindex_segdata=7;

% defining unique segment ids over all years for each highway/travelway
uniqueid =
segdata(:,highwayname_index_segdata)*10000000000+segdata(:,travelwayid_
index_segdata)*10000+segdata(:,27);
segdata(:,28) = uniqueid;
% the matrix that will include the final monthly count data
monthly_count_noremainid=[];
% separating crash/segment data for unique highway/travelway/year
unique_highway=unique(crashdata(:, highwayname_index_crashdata));
for highway = 1:length(unique_highway)
```

www.manaraa.com

```
index1=crashdata(:,highwayname_index_crashdata)==unique_highway(highway
);

index2=segdata(:,highwayname_index_segdata)==unique_highway(highway);
    highway_crashdata=crashdata(index1,:);
    highway_segdata=segdata(index2,:);


unique_travelway=unique(highway_crashdata(:,travelwayid_index_crashdata
));
    for travelway = 1:length(unique_travelway)
        index1=highway_crashdata(:,travelwayid_index_crashdata)==
unique_travelway(travelway);
        index2=highway_segdata(:,travelwayid_index_segdata)==
unique_travelway(travelway);
        travelway_highway_crashdata=highway_crashdata(index1,:);
        travelway_highway_segdata=highway_segdata(index2,:);


unique_year=unique(travelway_highway_crashdata(:,yearindex_crashdata));
        for year = 1: length(unique_year)
            if isempty(travelway_highway_segdata)==1;  break; end

index1=travelway_highway_crashdata(:,yearindex_crashdata)==unique_year(
year);

index2=travelway_highway_segdata(:,yearindex_segdata)==unique_year(year
);

year_travelway_highway_crashdata=travelway_highway_crashdata(index1,:);

year_travelway_highway_segdata=travelway_highway_segdata(index2,:);

% separating only crash data for unique
% highway/travelway/year/month

unique_month=unique(year_travelway_highway_crashdata(:,monthindex_crash
data));
            for i = 1:length(unique_month)

index1=year_travelway_highway_crashdata(:,monthindex_crashdata)==unique
_month(i);

month_year_travelway_highway_crashdata=year_travelway_highway_crashdata
(index1,:);

% separating crash data within unique
% highway/travelway/year/month for unique segments from the
% segdata separated within unique highway/travelway/year
                for segment=1:size(year_travelway_highway_segdata,1)

index=month_year_travelway_highway_crashdata(:,logindex_crashdata)>=
year_travelway_highway_segdata(segment,beglogindex_segdata) &...
```

www.manaraa.com

```
month_year_travelway_highway_crashdata(:,logindex_crashdata)<
year_travelway_highway_segdata(segment,endlogindex_segdata);

segment_month_year_travelway_highway_crashdata=month_year_travelway_hig
hway_crashdata(index,:);

segment_month_year_travelway_highway_count=size(segment_month_year_trav
elway_highway_crashdata,1);

% creating monthly count: column 29 is month and column 30 is the count
% for a highway-travelway-segment-year

monthly_count_segment=[year_travelway_highway_segdata(segment,:), i,
segment_month_year_travelway_highway_count];

monthly_count_noremainid=vertcat(monthly_count_noremainid,
monthly_count_segment);
                end
            end
        end
        t_stop=toc;
        hr=floor(t_stop/3600); mod1=mod(t_stop,3600);
        minut=floor(mod1/60); mod2=mod(t_stop,60);
        secon=floor(mod2);
        msg =strcat('   Counted monthly for highway-',
num2str(unique_highway(highway)), '  travelway-', ...
            num2str(unique_travelway(travelway)), '   Elapsed Time:',
num2str(hr),':', num2str(minut),':',num2str(secon));
        disp (msg)
    end
end

% separating unique segments from the monthly count data created above
unique_segment_monthly_count=unique(monthly_count_noremainid(:,28));
% the matrix that will include the final quarterly count data
quarterly_count_noremainid=[];
for segment=1:length(unique_segment_monthly_count)

index=monthly_count_noremainid(:,28)==unique_segment_monthly_count(segm
ent);
    uniquesegment_monthlydata=monthly_count_noremainid(index,:);

% creating quarterly count from the monthly count for each unique seg.
    quarter=0;
    for i=3:3:size(uniquesegment_monthlydata,1)
        if i+2 > size(uniquesegment_monthlydata,1); break; end
        uniquesegment_quarterdata=uniquesegment_monthlydata(i:i+2,:);
        quarter_count=sum(uniquesegment_quarterdata(:,30));
        quarter=quarter+1;

uniquesegment_quarterlydata=[uniquesegment_quarterdata(1,1:28),quarter,
quarter_count];

quarterly_count_noremainid=vertcat(quarterly_count_noremainid,uniqueseg
ment_quarterlydata);
```

www.manaraa.com

```
        if quarter == 4; quarter=0; end
        t_stop=toc;
        hr=floor(t_stop/3600); mod1=mod(t_stop,3600);
        minut=floor(mod1/60); mod2=mod(t_stop,60);
        secon=floor(mod2);
        msg =strcat('   Finishied counting quarterly for segment-',
num2str(segment),...
            ' th unique segment out of-',
num2str(length(unique_segment_monthly_count)),...
            '       Elapsed Time:', num2str(hr),':',...
            num2str(minut),':',num2str(secon));
        disp (msg)
    end
end

% separating unique segments from the monthly count data created above
% the matrix that will include the final yearly count data
yearly_count_noremainid=[];
for segment=1:length(unique_segment_monthly_count)

index=monthly_count_noremainid(:,28)==unique_segment_monthly_count(segm
ent);
    uniquesegment_monthlydata=monthly_count_noremainid(index,:);

    % creating yearly count from the monthly count for each unique seg.

unique_year_uniquesegment_monthlydata=unique(uniquesegment_monthlydata(
:,yearindex_segdata));
    for i=1: length(unique_year_uniquesegment_monthlydata)
        year=unique_year_uniquesegment_monthlydata(i);
        index=uniquesegment_monthlydata(:,yearindex_segdata)==year;
        uniquesegment_yeardata=uniquesegment_monthlydata(index,:);
        year_count=sum(uniquesegment_yeardata(:,30));

uniquesegment_yearlydata=[uniquesegment_yeardata(1,1:28),year,year_coun
t];
        yearly_count_noremainid=vertcat(yearly_count_noremainid,
uniquesegment_yearlydata);
        t_stop=toc;
        hr=floor(t_stop/3600); mod1=mod(t_stop,3600);
        minut=floor(mod1/60); mod2=mod(t_stop,60);
        secon=floor(mod2);
        msg =strcat('   Finishied counting Yearly for segment-',
num2str(segment),...
            ' th unique segment out of-',
num2str(length(unique_segment_monthly_count)),...
            '       Elapsed Time:', num2str(hr),':',...
            num2str(minut),':',num2str(secon));
        disp (msg)
    end
end

% the matrix that will include the final monthly, quarterly, and yearly
% count data plus the number of months, quarters, years remained
% unchanged
monthly_count=[]; quarterly_count=[]; yearly_count=[];
```

www.manaraa.com

```
for segment=1:length(unique_segment_monthly_count)

index1=monthly_count_noremainid(:,28)==unique_segment_monthly_count(seg
ment);
    uniquesegment_monthlydata=monthly_count_noremainid(index1,:);
    uniquesegment_monthlydata(:,31)=size(uniquesegment_monthlydata,1);
    monthly_count=vertcat(monthly_count,uniquesegment_monthlydata);


index2=quarterly_count_noremainid(:,28)==unique_segment_monthly_count(s
egment);
    uniquesegment_quarterlydata=quarterly_count_noremainid(index2,:);

uniquesegment_quarterlydata(:,31)=size(uniquesegment_quarterlydata,1);

quarterly_count=vertcat(quarterly_count,uniquesegment_quarterlydata);


index3=yearly_count_noremainid(:,28)==unique_segment_monthly_count(segm
ent);
    uniquesegment_yearlydata=yearly_count_noremainid(index3,:);
    uniquesegment_yearlydata(:,31)=size(uniquesegment_yearlydata,1);
    yearly_count=vertcat(yearly_count,uniquesegment_yearlydata);
end

%% ====save all the count files into a mat database=====
SAVEPATH=strcat(pwd,filesep,'matlab');
if ( ~isdir(SAVEPATH))
    mkdir(SAVEPATH);
end
SAVEFILENAME=strcat(SAVEPATH,filesep,'countdata_new.mat');
disp(['The count data was saved to: ',SAVEFILENAME]);
save(SAVEFILENAME, 'yearly_count', 'monthly_count', 'quarterly_count');

%% ====save all the count files into a three txt files for SAS=====
fileID = fopen('yearlycount_new.txt','w');
fprintf(fileID, '%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g\n',yearly_count');
fclose(fileID);

fileID = fopen('quarterlycount_new.txt','w');
fprintf(fileID, '%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g\n',quarterly_count');
fclose(fileID);

fileID = fopen('monthlycount_new.txt','w');
fprintf(fileID, '%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g\n',monthly_count');
fclose(fileID);

%% ====produce a sound after the run time is done=====
t1=1/10000:1/10000:1;
si=(sin(2*pi*500*t1));
sound(si,40000);
clear t1 si;
```

www.manaraa.com

APPENDIX B.

SAS CODES FOR MODELING CRASH FREQUENCY

www.manaraa.com

```
options ls=120 formdlim='-' nodate nonumber;
dm "out;clear;log;clear;";

*=====================================================;
data monthlycount;
set Seasonal.monthly;
where       sl > 50 and
            nolanes > 1 and
            year > 2001 and year < 2012
            ;
n=count;
LnAADT = log(AADT);
LnLength = log(segmentlength);
uniqueid = highway*10000000000+travelway*10000+id;
*-------------------------;
urban_2     = 0; urban_3 = 0; urban_3p = 0;
rural_2     = 0; rural_3 = 0;
*-------------------------;
urban_55 = 0; urban_60 = 0; urban_65 = 0; urban_70 = 0;
                    rural_60 = 0; rural_65 = 0; rural_70 = 0;
*-------------------------;
urban_2_55  = 0; urban_3_55   = 0; urban_3p_55 = 0;
urban_2_60  = 0; urban_3_60   = 0; urban_3p_60 = 0;
urban_2_65  = 0; urban_3_65   = 0; urban_3p_65 = 0;
urban_2_70  = 0; urban_3_70   = 0; urban_3p_70 = 0;
rural_2_60  = 0; rural_3_60   = 0;
rural_2_65  = 0; rural_3_65   = 0;
rural_2_70  = 0; rural_3_70   = 0;
*-------------------------;
lanewidthdummy    = 0;
*-------------------------;
psrclasslow = 0; psrclassmed = 0; psrclasshigh = 0;
*-------------------------;
nolanes2 = 0; nolanes3 = 0; nolanes3p = 0;
*-------------------------;
sl55 = 0; sl60 = 0; sl65 = 0; sl70 = 0;
*-------------------------;
spring_dummy = 0;
fall_dummy   = 0;
summer_dummy = 0;
winter_dummy = 0;
*-------------------------;
spring_2    = 0; spring_3     = 0; spring_3p    = 0;
summer_2    = 0; summer_3     = 0; summer_3p    = 0;
fall_2           = 0; fall_3       = 0; fall_3p      = 0;
winter_2    = 0; winter_3     = 0; winter_3p    = 0;
*-------------------------;
spring_55   = 0; summer_55    = 0; fall_55      = 0; winter_55    = 0;
spring_60   = 0; summer_60    = 0; fall_60      = 0; winter_60    = 0;
spring_65   = 0; summer_65    = 0; fall_65      = 0; winter_65    = 0;
spring_70   = 0; summer_70    = 0; fall_70      = 0; winter_70    = 0;
*-------------------------;
spring_2_55 = 0; spring_3_55 = 0; spring_3p_55 = 0;
spring_2_60 = 0; spring_3_60 = 0; spring_3p_60 = 0;
spring_2_65 = 0; spring_3_65 = 0; spring_3p_65 = 0;
spring_2_70 = 0; spring_3_70 = 0; spring_3p_70 = 0;
```

www.manaraa.com

```
summer_2_55 = 0; summer_3_55 = 0; summer_3p_55 = 0;
summer_2_60 = 0; summer_3_60 = 0; summer_3p_60 = 0;
summer_2_65 = 0; summer_3_65 = 0; summer_3p_65 = 0;
summer_2_70 = 0; summer_3_70 = 0; summer_3p_70 = 0;
fall_2_55   = 0; fall_3_55     = 0; fall_3p_55   = 0;
fall_2_60   = 0; fall_3_60     = 0; fall_3p_60   = 0;
fall_2_65   = 0; fall_3_65     = 0; fall_3p_65   = 0;
fall_2_70   = 0; fall_3_70     = 0; fall_3p_70   = 0;
winter_2_55 = 0; winter_3_55 = 0; winter_3p_55 = 0;
winter_2_60 = 0; winter_3_60 = 0; winter_3p_60 = 0;
winter_2_65 = 0; winter_3_65 = 0; winter_3p_65 = 0;
winter_2_70 = 0; winter_3_70 = 0; winter_3p_70 = 0;
*-------------------------;
fall_rural_2_60 = 0; fall_rural_2_65 = 0; fall_rural_2_70 = 0;
fall_rural_3_60 = 0; fall_rural_3_65 = 0; fall_rural_3_70 = 0;
fall_urban_2_55 = 0; fall_urban_2_60 = 0; fall_urban_2_65 = 0;
fall_urban_2_70 = 0;
fall_urban_3_55 = 0; fall_urban_3_60 = 0; fall_urban_3_65 = 0;
fall_urban_3_70 = 0;
fall_urban_3p_55 = 0; fall_urban_3p_60 = 0; fall_urban_3p_65 = 0;
fall_urban_3p_70 = 0;
spring_rural_2_60 = 0; spring_rural_2_65 = 0; spring_rural_2_70 = 0;
spring_rural_3_60 = 0; spring_rural_3_65 = 0; spring_rural_3_70 = 0;
spring_urban_2_55 = 0; spring_urban_2_60 = 0; spring_urban_2_65 = 0;
spring_urban_2_70 = 0;
spring_urban_3_55 = 0; spring_urban_3_60 = 0; spring_urban_3_65 = 0;
spring_urban_3_70 = 0;
spring_urban_3p_55 = 0; spring_urban_3p_60 = 0; spring_urban_3p_65 = 0;
spring_urban_3p_70 = 0;
summer_rural_2_60 = 0; summer_rural_2_65 = 0; summer_rural_2_70 = 0;
summer_rural_3_60 = 0; summer_rural_3_65 = 0; summer_rural_3_70 = 0;
summer_urban_2_55 = 0; summer_urban_2_60 = 0; summer_urban_2_65 = 0;
summer_urban_2_70 = 0;
summer_urban_3_55 = 0; summer_urban_3_60 = 0; summer_urban_3_65 = 0;
summer_urban_3_70 = 0;
summer_urban_3p_55 = 0; summer_urban_3p_60 = 0; summer_urban_3p_65 = 0;
summer_urban_3p_70 = 0;
winter_rural_2_60 = 0; winter_rural_2_65 = 0; winter_rural_2_70 = 0;
winter_rural_3_60 = 0; winter_rural_3_65 = 0; winter_rural_3_70 = 0;
winter_urban_2_55 = 0; winter_urban_2_60 = 0; winter_urban_2_65 = 0;
winter_urban_2_70 = 0;
winter_urban_3_55 = 0; winter_urban_3_60 = 0; winter_urban_3_65 = 0;
winter_urban_3_70 = 0;
winter_urban_3p_55 = 0; winter_urban_3p_60 = 0; winter_urban_3p_65 = 0;
winter_urban_3p_70= 0;
*-------------------------;
transition = 0;
spring_transition = 0;
summer_transition = 0;
*fall_transition    = 0;
winter_transition = 0;
*-------------------------;

*####################################################;
*####################################################;
*####################################################;
```

www.manaraa.com

```
*####################################################;
proc format;
      value area  1='Urban'
                  0='Rural';
run;

data count;
set monthlycount;
*--------------------------;
if year > 2001 then overallmonth = (year-2002)*12+month;
if year > 2004 then     transition = (overallmonth-36)/84;
* the 35 is the number of months for 2002- November 2004;
*--------------------------;
if month = 3  or month = 4  or month = 5  then do spring_dummy = 1;
season = 1; end;
if month = 6  or month = 7  or month = 8  then do summer_dummy = 1;
season = 2; end;
if month = 9  or month = 10 or month = 11 then do fall_dummy   = 1;
season = 3; end;
if month = 12 or month = 1  or month = 2  then do winter_dummy = 1;
season = 4; end;
*--------------------------;
if spring_dummy = 1 and nolanes = 2 then spring_2 = 1;
if summer_dummy = 1 and nolanes = 2 then summer_2 = 1;
if fall_dummy   = 1 and nolanes = 2 then fall_2   = 1;
if winter_dummy = 1 and nolanes = 2 then winter_2 = 1;
if spring_dummy = 1 and nolanes = 3 then spring_3 = 1;
if summer_dummy = 1 and nolanes = 3 then summer_3 = 1;
if fall_dummy   = 1 and nolanes = 3 then fall_3   = 1;
if winter_dummy = 1 and nolanes = 3 then winter_3 = 1;
if spring_dummy = 1 and nolanes > 3 then spring_3p= 1;
if summer_dummy = 1 and nolanes > 3 then summer_3p= 1;
if fall_dummy   = 1 and nolanes > 3 then fall_3p       = 1;
if winter_dummy = 1 and nolanes > 3 then winter_3p= 1;
*--------------------------;
if spring_dummy = 1 and sl = 55 then spring_55 = 1;
if summer_dummy = 1 and sl = 55 then summer_55 = 1;
if fall_dummy   = 1 and sl = 55 then fall_55   = 1;
if winter_dummy = 1 and sl = 55 then winter_55 = 1;
if spring_dummy = 1 and sl = 60 then spring_60 = 1;
if summer_dummy = 1 and sl = 60 then summer_60 = 1;
if fall_dummy   = 1 and sl = 60 then fall_60   = 1;
if winter_dummy = 1 and sl = 60 then winter_60 = 1;
if spring_dummy = 1 and sl = 65 then spring_65 = 1;
if summer_dummy = 1 and sl = 65 then summer_65 = 1;
if fall_dummy   = 1 and sl = 65 then fall_65   = 1;
if winter_dummy = 1 and sl = 65 then winter_65 = 1;
if spring_dummy = 1 and sl = 70 then spring_70 = 1;
if summer_dummy = 1 and sl = 70 then summer_70 = 1;
if fall_dummy   = 1 and sl = 70 then fall_70   = 1;
if winter_dummy = 1 and sl = 70 then winter_70 = 1;
*--------------------------;
if nolanes = 2 then nolanes2 = 1;
if nolanes = 3 then nolanes3 = 1;
if nolanes > 3 then nolanes3p = 1;
*--------------------------;
```

www.manaraa.com

```
if sl = 55 then sl55 = 1;
if sl = 60 then sl60 = 1;
if sl = 65 then sl65 = 1;
if sl = 70 then sl70 = 1;
*-------------------------;
if area = 1 and nolanes = 2 then urban_2 = 1;
if area = 1 and nolanes = 3 then urban_3 = 1;
if area = 1 and nolanes > 3 then urban_3p= 1;
if area = 0 and nolanes = 2 then rural_2 = 1;
if area = 0 and nolanes = 3 then rural_3 = 1;
*-------------------------;
if area = 1 and sl = 55 then urban_55     = 1;
if area = 1 and sl = 60 then urban_60     = 1;
if area = 1 and sl = 65 then urban_65     = 1;
if area = 1 and sl = 70 then urban_70     = 1;
if area = 0 and sl = 60 then rural_60     = 1;
if area = 0 and sl = 65 then rural_65     = 1;
if area = 0 and sl = 70 then rural_70     = 1;
*-------------------------;
if area = 1 and nolanes = 2 and sl = 55 then urban_2_55 = 1;
if area = 1 and nolanes = 2 and sl = 55 then
            select (season);
                  when (1) spring_urban_2_55    =1;
                  when (2) summer_urban_2_55    =1;
                  when (3) fall_urban_2_55      =1;
                  when (4) winter_urban_2_55    =1;
            end;
if area = 1 and nolanes = 2 and sl = 60 then urban_2_60 = 1;
if area = 1 and nolanes = 2 and sl = 60 then
            select (season);
                  when (1) spring_urban_2_60    =1;
                  when (2) summer_urban_2_60    =1;
                  when (3) fall_urban_2_60      =1;
                  when (4) winter_urban_2_60    =1;
            end;
if area = 1 and nolanes = 2 and sl = 65 then urban_2_65 = 1;
if area = 1 and nolanes = 2 and sl = 65 then
            select (season);
                  when (1) spring_urban_2_65    =1;
                  when (2) summer_urban_2_65    =1;
                  when (3) fall_urban_2_65      =1;
                  when (4) winter_urban_2_65    =1;
            end;
if area = 1 and nolanes = 2 and sl = 70 then urban_2_70     = 1;
if area = 1 and nolanes = 2 and sl = 70 then
            select (season);
                  when (1) spring_urban_2_70    =1;
                  when (2) summer_urban_2_70    =1;
                  when (3) fall_urban_2_70      =1;
                  when (4) winter_urban_2_70    =1;
            end;
if area = 1 and nolanes = 3 and sl = 55 then urban_3_55     = 1;
if area = 1 and nolanes = 3 and sl = 55 then
            select (season);
                  when (1) spring_urban_3_55    =1;
                  when (2) summer_urban_3_55    =1;
```

www.manaraa.com

```
                when (3) fall_urban_3_55      =1;
                when (4) winter_urban_3_55    =1;
            end;
if area = 1 and nolanes = 3 and sl = 60 then urban_3_60     = 1;
if area = 1 and nolanes = 3 and sl = 60 then
            select (season);
                when (1) spring_urban_3_60    =1;
                when (2) summer_urban_3_60    =1;
                when (3) fall_urban_3_60      =1;
                when (4) winter_urban_3_60    =1;
            end;
if area = 1 and nolanes = 3 and sl = 65 then urban_3_65     = 1;
if area = 1 and nolanes = 3 and sl = 65 then
            select (season);
                when (1) spring_urban_3_65    =1;
                when (2) summer_urban_3_65    =1;
                when (3) fall_urban_3_65      =1;
                when (4) winter_urban_3_65    =1;
            end;
if area = 1 and nolanes = 3 and sl = 70 then urban_3_70     = 1;
if area = 1 and nolanes = 3 and sl = 70 then
            select (season);
                when (1) spring_urban_3_70    =1;
                when (2) summer_urban_3_70    =1;
                when (3) fall_urban_3_70      =1;
                when (4) winter_urban_3_70    =1;
            end;
if area = 1 and nolanes > 3 and sl = 55 then urban_3p_55 = 1;
if area = 1 and nolanes > 3 and sl = 55 then
            select (season);
                when (1) spring_urban_3p_55   =1;
                when (2) summer_urban_3p_55   =1;
                when (3) fall_urban_3p_55     =1;
                when (4) winter_urban_3p_55   =1;
            end;
if area = 1 and nolanes > 3 and sl = 60 then urban_3p_60 = 1;
if area = 1 and nolanes > 3 and sl = 60 then
            select (season);
                when (1) spring_urban_3p_60   =1;
                when (2) summer_urban_3p_60   =1;
                when (3) fall_urban_3p_60     =1;
                when (4) winter_urban_3p_60   =1;
            end;
if area = 1 and nolanes > 3 and sl = 65 then urban_3p_65 = 1;
if area = 1 and nolanes > 3 and sl = 65 then
            select (season);
                when (1) spring_urban_3p_65   =1;
                when (2) summer_urban_3p_65   =1;
                when (3) fall_urban_3p_65     =1;
                when (4) winter_urban_3p_65   =1;
            end;
if area = 1 and nolanes > 3 and sl = 70 then urban_3p_70 = 1;
if area = 1 and nolanes > 3 and sl = 70 then
            select (season);
                when (1) spring_urban_3p_70   =1;
                when (2) summer_urban_3p_70   =1;
```

www.manaraa.com

```
                when (3) fall_urban_3p_70     =1;
                when (4) winter_urban_3p_70   =1;
            end;
if area = 0 and nolanes = 2 and sl = 60 then rural_2_60     = 1;
if area = 0 and nolanes = 2 and sl = 60 then
            select (season);
                when (1) spring_rural_2_60    =1;
                when (2) summer_rural_2_60    =1;
                when (3) fall_rural_2_60      =1;
                when (4) winter_rural_2_60    =1;
            end;
if area = 0 and nolanes = 2 and sl = 65 then rural_2_65     = 1;
if area = 0 and nolanes = 2 and sl = 65 then
            select (season);
                when (1) spring_rural_2_65    =1;
                when (2) summer_rural_2_65    =1;
                when (3) fall_rural_2_65      =1;
                when (4) winter_rural_2_65    =1;
            end;
if area = 0 and nolanes = 2 and sl = 70 then rural_2_70     = 1;
if area = 0 and nolanes = 2 and sl = 70 then
            select (season);
                when (1) spring_rural_2_70    =1;
                when (2) summer_rural_2_70    =1;
                when (3) fall_rural_2_70      =1;
                when (4) winter_rural_2_70    =1;
            end;
if area = 0 and nolanes = 3 and sl = 60 then rural_3_60     = 1;
if area = 0 and nolanes = 3 and sl = 60 then
            select (season);
                when (1) spring_rural_3_60    =1;
                when (2) summer_rural_3_60    =1;
                when (3) fall_rural_3_60      =1;
                when (4) winter_rural_3_60    =1;
            end;
if area = 0 and nolanes = 3 and sl = 65 then rural_3_65     = 1;
if area = 0 and nolanes = 3 and sl = 65 then
            select (season);
                when (1) spring_rural_3_65    =1;
                when (2) summer_rural_3_65    =1;
                when (3) fall_rural_3_65      =1;
                when (4) winter_rural_3_65    =1;
            end;
if area = 0 and nolanes = 3 and sl = 70 then rural_3_70     = 1;
if area = 0 and nolanes = 3 and sl = 70 then
            select (season);
                when (1) spring_rural_3_70    =1;
                when (2) summer_rural_3_70    =1;
                when (3) fall_rural_3_70      =1;
                when (4) winter_rural_3_70    =1;
            end;
*-------------------------;
if lanewidth ne 12 then lanewidthdummy = 1;
*-------------------------;
areadt                  = area * lnAADT;
areacommercial    = area * percentcommercial;
```

www.manaraa.com

```
areawidth          = area * lanewidth;
areapsr                 = area * psr;
areacongestion     = area * congestionindex;
areawidthdummy     = area * lanewidthdummy;
areatransition   = area * transition;
areaseason         = area * season;
*-------------------------;
urban_2dt    = urban_2  * Lnaadt;
urban_3dt    = urban_3  * Lnaadt;
urban_3pdt   = urban_3p * Lnaadt;
*-------------------------;
if     psr < (32.498298-.3*2.63746) then do psrclasslow = 1; psrclass=1;
end;
if     psr > (32.498298-.3*2.63746) and psr < (32.498298+.3*2.63746)
then do psrclassmed = 1; psrclass=2; end;
if     psr > (32.498298+.3*2.63746) then do psrclasshigh = 1;
psrclass=3; end;
*-------------------------;
* Creating interactions with seasonal dummy variables;
spring_area                    = spring_dummy*area;
summer_area                    = summer_dummy*area;
winter_area                    = winter_dummy*area;
spring_nolanes                 = spring_dummy*nolanes;
summer_nolanes                 = summer_dummy*nolanes;
winter_nolanes                 = winter_dummy*nolanes;
spring_lanewidth          = spring_dummy*lanewidth;
summer_lanewidth          = summer_dummy*lanewidth;
winter_lanewidth          = winter_dummy*lanewidth;
spring_shoulderwidth      = spring_dummy*shoulderwidth;
summer_shoulderwidth      = summer_dummy*shoulderwidth;
winter_shoulderwidth      = winter_dummy*shoulderwidth;
spring_lnAADT                  = spring_dummy*lnAADT;
summer_lnAADT                  = summer_dummy*lnAADT;
winter_lnAADT                  = winter_dummy*lnAADT;
spring_SL                      = spring_dummy*SL;
summer_SL                      = summer_dummy*SL;
winter_SL                      = winter_dummy*SL;
spring_congestionindex    = spring_dummy*congestionindex;
summer_congestionindex    = summer_dummy*congestionindex;
winter_congestionindex    = winter_dummy*congestionindex;
spring_PSR                     = spring_dummy*PSR;
summer_PSR                     = summer_dummy*PSR;
winter_PSR                     = winter_dummy*PSR;
spring_percentcommercial  = spring_dummy*percentcommercial;
summer_percentcommercial  = summer_dummy*percentcommercial;
winter_percentcommercial  = winter_dummy*percentcommercial;
spring_transition         = spring_dummy*transition;
summer_transition         = summer_dummy*transition;
winter_transition         = winter_dummy*transition;
*-------------------------;
/*
if spring_dummy = 1 and nolanes = 2 and sl = 55 then spring_2_55 = 1;
if spring_dummy = 1 and nolanes = 3 and sl = 55 then spring_3_55 = 1;
if spring_dummy = 1 and nolanes > 3 and sl = 55 then spring_3p_55= 1;
if spring_dummy = 1 and nolanes = 2 and sl = 60 then spring_2_60 = 1;
if spring_dummy = 1 and nolanes = 3 and sl = 60 then spring_3_60 = 1;
```

www.manaraa.com

```
if spring_dummy = 1 and nolanes > 3 and sl = 60 then spring_3p_60= 1;
if spring_dummy = 1 and nolanes = 2 and sl = 65 then spring_2_65 = 1;
if spring_dummy = 1 and nolanes = 3 and sl = 65 then spring_3_65 = 1;
if spring_dummy = 1 and nolanes > 3 and sl = 65 then spring_3p_65= 1;
if spring_dummy = 1 and nolanes = 2 and sl = 70 then spring_2_70 = 1;
if spring_dummy = 1 and nolanes = 3 and sl = 70 then spring_3_70 = 1;
if spring_dummy = 1 and nolanes > 3 and sl = 70 then spring_3p_70= 1;

if summer_dummy = 1 and nolanes = 2 and sl = 55 then summer_2_55 = 1;
if summer_dummy = 1 and nolanes = 3 and sl = 55 then summer_3_55 = 1;
if summer_dummy = 1 and nolanes > 3 and sl = 55 then summer_3p_55= 1;
if summer_dummy = 1 and nolanes = 2 and sl = 60 then summer_2_60 = 1;
if summer_dummy = 1 and nolanes = 3 and sl = 60 then summer_3_60 = 1;
if summer_dummy = 1 and nolanes > 3 and sl = 60 then summer_3p_60= 1;
if summer_dummy = 1 and nolanes = 2 and sl = 65 then summer_2_65 = 1;
if summer_dummy = 1 and nolanes = 3 and sl = 65 then summer_3_65 = 1;
if summer_dummy = 1 and nolanes > 3 and sl = 65 then summer_3p_65= 1;
if summer_dummy = 1 and nolanes = 2 and sl = 70 then summer_2_70 = 1;
if summer_dummy = 1 and nolanes = 3 and sl = 70 then summer_3_70 = 1;
if summer_dummy = 1 and nolanes > 3 and sl = 70 then summer_3p_70= 1;

if fall_dummy   = 1 and nolanes = 2 and sl = 55 then fall_2_55    = 1;
if fall_dummy   = 1 and nolanes = 3 and sl = 55 then fall_3_55    = 1;
if fall_dummy   = 1 and nolanes > 3 and sl = 55 then fall_3p_55   = 1;
if fall_dummy   = 1 and nolanes = 2 and sl = 60 then fall_2_60    = 1;
if fall_dummy   = 1 and nolanes = 3 and sl = 60 then fall_3_60    = 1;
if fall_dummy   = 1 and nolanes > 3 and sl = 60 then fall_3p_60   = 1;
if fall_dummy   = 1 and nolanes = 2 and sl = 65 then fall_2_65    = 1;
if fall_dummy   = 1 and nolanes = 3 and sl = 65 then fall_3_65    = 1;
if fall_dummy   = 1 and nolanes > 3 and sl = 65 then fall_3p_65   = 1;
if fall_dummy   = 1 and nolanes = 2 and sl = 70 then fall_2_70    = 1;
if fall_dummy   = 1 and nolanes = 3 and sl = 70 then fall_3_70    = 1;
if fall_dummy   = 1 and nolanes > 3 and sl = 70 then fall_3p_70   = 1;
*/
if winter_dummy = 1 and area = 1 and nolanes = 2 and sl = 55 then
winter_urban_2_55 = 1;
if winter_dummy = 1 and area = 1 and nolanes = 3 and sl = 55 then
winter_urban_3_55 = 1;
if winter_dummy = 1 and area = 1 and nolanes > 3 and sl = 55 then
winter_urban_3p_55= 1;
if winter_dummy = 1 and area = 1 and nolanes = 2 and sl = 60 then
winter_urban_2_60 = 1;
if winter_dummy = 1 and area = 1 and nolanes = 3 and sl = 60 then
winter_urban_3_60 = 1;
if winter_dummy = 1 and area = 1 and nolanes > 3 and sl = 60 then
winter_urban_3p_60= 1;
if winter_dummy = 1 and area = 1 and nolanes = 2 and sl = 65 then
winter_urban_2_65 = 1;
if winter_dummy = 1 and area = 1 and nolanes = 3 and sl = 65 then
winter_urban_3_65 = 1;
if winter_dummy = 1 and area = 1 and nolanes > 3 and sl = 65 then
winter_urban_3p_65= 1;
if winter_dummy = 1 and area = 1 and nolanes = 2 and sl = 70 then
winter_urban_2_70 = 1;
if winter_dummy = 1 and area = 1 and nolanes = 3 and sl = 70 then
winter_urban_3_70 = 1;
```

www.manaraa.com

```
if winter_dummy = 1 and area = 1 and nolanes > 3 and sl = 70 then
winter_urban_3p_70= 1;
*-------------------------;
if spring_dummy = 1 and area = 1 and nolanes = 2 and sl = 55 then
spring_urban_2_55 = 1;
if spring_dummy = 1 and area = 1 and nolanes = 3 and sl = 55 then
spring_urban_3_55 = 1;
if spring_dummy = 1 and area = 1 and nolanes > 3 and sl = 55 then
spring_urban_3p_55= 1;
if spring_dummy = 1 and area = 1 and nolanes = 2 and sl = 60 then
spring_urban_2_60 = 1;
if spring_dummy = 1 and area = 1 and nolanes = 3 and sl = 60 then
spring_urban_3_60 = 1;
if spring_dummy = 1 and area = 1 and (nolanes > 3 or nolanes=2) and (sl
= 60 or sl=70) then spring_urban_3p_60= 1;
if spring_dummy = 1 and area = 1 and nolanes = 2 and sl = 65 then
spring_urban_2_65 = 1;
if spring_dummy = 1 and area = 1 and nolanes = 3 and sl = 65 then
spring_urban_3_65 = 1;
if spring_dummy = 1 and area = 1 and nolanes > 3 and sl = 65 then
spring_urban_3p_65= 1;
if spring_dummy = 1 and area = 1 and nolanes = 2 and sl = 70 then
spring_urban_2_70 = 1;
if spring_dummy = 1 and area = 1 and nolanes = 3 and sl = 70 then
spring_urban_3_70 = 1;
if spring_dummy = 1 and area = 1 and nolanes > 3 and sl = 70 then
spring_urban_3p_70= 1;
*-------------------------;
dataset = "original";
run;

*#####################################################;
*#####################################################;
*###### This section is for the modeling which was done first to find
the sig. variables
/*********************************************************************
**********/
*GEE NB model. Each segment count is a repeated observation | With
Interaction;
proc sort data = count; by uniqueid year month; run;

proc genmod data = count;
where highway=29 or highway=35 or highway=44 or highway=49 or
highway=55 or highway=70;
class       uniqueid ;
model       n = lnAADT
                psr
                percentcommercial
                congestionindex
                transition
                areadt
                areacommercial
/*              lanewidthdummy Removed backward! */
/*              areapsr Removed backward! */
/*              areatransition Removed backward! */
                areacongestion
```

www.manaraa.com

```
                urban_2_55
                urban_2_60
                urban_2_65
                urban_2_70
                urban_3_55
                urban_3_60
                urban_3_65
                urban_3_70
                urban_3p_55
                urban_3p_60
                urban_3p_65
                urban_3p_70
                rural_2_60
                rural_2_65
/*              rural_2_70 base for area_nolane_sl */

/*              spring_area Removed backward! */
                spring_dummy
                spring_lnAADT
/*              spring_percentcommercial Removed backward! */
                spring_PSR
/*              spring_lanewidth Removed backward! */
/*              spring_rural_2_60 Removed backward! */
/*              spring_rural_2_65 Removed backward! */
/*              spring_rural_2_70 Removed backward! */
/*              spring_transition Removed backward! */
/*              spring_urban_2_55 Removed backward! */
/*              spring_urban_2_60 Removed backward! */
/*              spring_urban_2_65 Removed backward! */
                spring_urban_2_70
/*              spring_urban_3_55 Removed backward! */
/*              spring_urban_3_60 Removed backward! */
/*              spring_urban_3_65 Removed backward! */
/*              spring_urban_3_70 Removed backward! */
/*              spring_urban_3p_55 Removed backward! */
/*              spring_urban_3p_60 Removed backward! */
/*              spring_urban_3p_65 Removed backward! */
/*              spring_urban_3p_70 base for spring_area_nolane_sl */

/*              summer_area Removed backward! */
                summer_dummy
/*              summer_lnAADT Removed backward! */
                summer_percentcommercial
                summer_PSR
/*              summer_lanewidth Removed backward! */
/*              summer_rural_2_60 Removed backward! */
/*              summer_rural_2_65 Removed backward! */
/*              summer_rural_2_70 Removed backward! */
                summer_transition
/*              summer_urban_2_55 Removed backward! */
/*              summer_urban_2_60 Removed backward! */
/*              summer_urban_2_65 Removed backward! */
/*              summer_urban_2_70 Removed backward! */
                summer_urban_3_55
/*              summer_urban_3_60 Removed backward! */
/*              summer_urban_3_65 Removed backward! */
```

www.manaraa.com

```
/*                summer_urban_3_70 Removed backward! */
/*                summer_urban_3p_55 Removed backward! */
/*                summer_urban_3p_60 Removed backward! */
/*                summer_urban_3p_65 Removed backward! */
/*                summer_urban_3p_70 base for summer_area_nolane_sl */

/*                winter_area Removed backward! */
                  winter_dummy
                  winter_lnAADT
/*                winter_percentcommercial Removed backward! */
                  winter_PSR
/*                winter_lanewidth Removed backward! */
/*                winter_rural_2_60 Removed backward! */
/*                winter_rural_2_65 Removed backward! */
/*                winter_rural_2_70 Removed backward! */
                  winter_transition
                  winter_urban_2_55
/*                winter_urban_2_60 Removed backward! */
                  winter_urban_2_65
/*                winter_urban_2_70 Removed backward! */
                  winter_urban_3_55
                  winter_urban_3_60
                  winter_urban_3_65
                  winter_urban_3_70
                  winter_urban_3p_55
                  winter_urban_3p_60
                  winter_urban_3p_65
/*                winter_urban_3p_70 base for winter_area_nolane_sl */

             /offset= segmentLength d=nb;
repeated subject=uniqueid / type=ar;
assess var=(lnaadt)/ resample=1000;
run;

*#####################################################################;
*#####################################################################;
*###### This section is for finding the effect of seasons after
*###### the modeling is finalized. only sig. variables are used here;
*#####################################################################;
*#####################################################;
*#####################################################;
*#####################################################;
*#####################################################;

data countnew;
set count;
      n                     = .;
      lnAADT                = 0;
      psr                   = 0;
      percentcommercial     = 0;
      congestionindex       = 0;
      transition            = 0;
      areadt                = 0;
      areacommercial        = 0;
      lanewidthdummy        = 0;
      areapsr               = 0;
```

www.manaraa.com

```
        areatransition          = 0;
        areacongestion          = 0;

        urban_2_55              = 0;
        urban_2_60              = 0;
        urban_2_65              = 0;
        urban_2_70              = 0;
        urban_3_55              = 0;
        urban_3_60              = 0;
        urban_3_65              = 0;
        urban_3_70              = 0;
        urban_3p_55             = 0;
        urban_3p_60             = 0;
        urban_3p_65             = 0;
        urban_3p_70             = 0;
        rural_2_60              = 0;
        rural_2_65              = 0;
        rural_2_70              = 0;
/*
        spring_area             = 0;
        spring_dummy            = 0;
        spring_lnAADT           = 0;
        spring_percentcommercial = 0;
        spring_PSR              = 0;
        spring_lanewidth  = 0;
        spring_rural_2_60       = 0;
        spring_rural_2_65       = 0;
        spring_rural_2_70       = 0;
        spring_transition       = 0;
        spring_urban_2_55       = 0;
        spring_urban_2_60       = 0;
        spring_urban_2_65       = 0;
        spring_urban_2_70       = 0;
        spring_urban_3_55       = 0;
        spring_urban_3_60       = 0;
        spring_urban_3_65       = 0;
        spring_urban_3_70       = 0;
        spring_urban_3p_55      = 0;
        spring_urban_3p_60      = 0;
        spring_urban_3p_65      = 0;
        spring_urban_3p_70      = 0;
*/
        summer_area             = 0;
        summer_dummy            = 0;
        summer_lnAADT           = 0;
        summer_percentcommercial= 0;
        summer_PSR              = 0;
        summer_lanewidth        = 0;
        summer_rural_2_60       = 0;
        summer_rural_2_65       = 0;
        summer_rural_2_70       = 0;
        summer_transition       = 0;
        summer_urban_2_55       = 0;
        summer_urban_2_60       = 0;
        summer_urban_2_65       = 0;
        summer_urban_2_70       = 0;
```

www.manaraa.com

```
      summer_urban_3_55       = 0;
      summer_urban_3_60       = 0;
      summer_urban_3_65       = 0;
      summer_urban_3_70       = 0;
      summer_urban_3p_55      = 0;
      summer_urban_3p_60      = 0;
      summer_urban_3p_65      = 0;
      summer_urban_3p_70      = 0;

      winter_area             = 0;
      winter_dummy            = 0;
      winter_lnAADT           = 0;
      winter_percentcommercial= 0;
      winter_PSR              = 0;
      winter_lanewidth        = 0;
      winter_rural_2_60       = 0;
      winter_rural_2_65       = 0;
      winter_rural_2_70       = 0;
      winter_transition       = 0;
      winter_urban_2_55       = 0;
      winter_urban_2_60       = 0;
      winter_urban_2_65       = 0;
      winter_urban_2_70       = 0;
      winter_urban_3_55       = 0;
      winter_urban_3_60       = 0;
      winter_urban_3_65       = 0;
      winter_urban_3_70       = 0;
      winter_urban_3p_55      = 0;
      winter_urban_3p_60      = 0;
      winter_urban_3p_65      = 0;
      winter_urban_3p_70      = 0;

      dataset                 = "new";
run;
*#####################################################;
*#####################################################;
*=====================================================;
data countfinal;
set count countnew;
run;
*=====================================================;
proc genmod data = countfinal;
where highway=29 or highway=35 or highway=44 or highway=49 or
highway=55 or highway=70;
class       uniqueid ;
model       n = lnAADT
                psr
                percentcommercial
                congestionindex
                transition
                areadt
                areacommercial
                areacongestion
                urban_2_55
                urban_2_60
                urban_2_65
```

www.manaraa.com

```
                    urban_2_70
                    urban_3_55
                    urban_3_60
                    urban_3_65
                    urban_3_70
                    urban_3p_55
                    urban_3p_60
                    urban_3p_65
                    urban_3p_70
                    rural_2_60
                    rural_2_65

                    spring_dummy
                    spring_lnAADT
                    spring_PSR
                    spring_urban_2_70

                    summer_dummy
                    summer_percentcommercial
                    summer_PSR
                    summer_transition
                    summer_urban_3_55

                    winter_dummy
                    winter_lnAADT
                    winter_PSR
                    winter_transition
                    winter_urban_2_55
                    winter_urban_2_65
                    winter_urban_3_55
                    winter_urban_3_60
                    winter_urban_3_65
                    winter_urban_3_70
                    winter_urban_3p_55
                    winter_urban_3p_60
                    winter_urban_3p_65

             /offset= segmentLength d=nb;
repeated subject=uniqueid / type=ar;
output out=preddata lower=lowerb
                    upper=upperb
                    resraw=residuals
                    xbeta=linear_function_values
                    pred=predicted_values;
run;
************************************************;
proc print data = preddata (obs=10);
var uniqueid n  predicted_values residuals;
where summer_dummy = 1 and dataset="new";
run;

proc means data = preddata min max mean var sum;
var uniqueid n predicted_values;
where summer_dummy = 1 and dataset="new";
run;
```

www.manaraa.com

APPENDIX C.

DETAILS OF THE EXAMINATION FOR CONFOUNDING AND SUFFICIENCY OF OBSERVATIONS

www.manaraa.com

Examining Shoulderwidth

The shoulder width and PSR were examined for the number of observations in each class to verify the sufficiency for analysis. Table 1 shows this tabularization for the shoulder width and year. In each group there are two numbers presented in the table that shows the actual number of observation and the overall percentage for that group. For example there are 31 segments observed with shoulderwidth=4' during the year 2002 which is 0.46% of the overall number of observations. It can be observed that for all the shoulder widths except 10' there are not enough observations (as suggested by the reviewer to be at least 60). The histogram of the total number of observations is also presented in Figure 1).

Table 1 clearly shows that the distribution of shoulderwidth that is present in this data set eliminates the possibility of grouping this data into reasonable categories based on shoulderwidth because 86% of the observations have shoulderwidth=10. Categorizing the observations according to the mean (9.4814107) and sd (1.5621976) of this variable as suggested by the reviewer, will also not work as seen in Table 2. Using the 25^{th}, 50^{th}, and 75^{th} percentiles as the divider to group observations in sufficient numbers also will not resolve the problem (Table 3) as all three percentile values were one number (equal 10'). Therefore, it was decided to remove the variable shoulderwidth from the analysis noting that it was also found to be not statistically significant (p-value= 0.9596)in the old series of analyses.

www.manaraa.com

Table 1. Frequency and percentages for each class of shoulder width within each year

year	shoulderwidth														
	0	1	2	3	4	5	6	7	8	9	10	11	12	15	Total
2002	0	0	5	0	31	5	36	5	49	18	916	0	11	0	1076
	0	0	0.07	0	0.46	0.07	0.53	0.07	0.72	0.26	13.5	0	0.16	0	15.81
2003	4	0	6	0	36	5	46	7	47	17	891	0	14	0	1073
	0.06	0	0.09	0	0.53	0.07	0.68	0.1	0.69	0.25	13.1	0	0.21	0	15.77
2004	4	2	5	2	38	13	52	7	52	17	867	0	17	1	1077
	0.06	0.03	0.07	0.03	0.56	0.19	0.76	0.1	0.76	0.25	12.7	0	0.25	0	15.83
2005	3	2	3	0	37	13	44	4	44	16	771	0	17	1	955
	0.04	0.03	0.04	0	0.54	0.19	0.65	0.06	0.65	0.24	11.3	0	0.25	0	14.03
2006	0	1	0	1	3	0	5	4	5	12	210	0	4	0	245
	0	0.01	0	0.01	0.04	0	0.07	0.06	0.07	0.18	3.09	0	0.06	0	3.6
2007	0	1	0	2	7	4	9	4	15	12	400	0	7	0	461
	0	0.01	0	0.03	0.1	0.06	0.13	0.06	0.22	0.18	5.88	0	0.1	0	6.77
2008	0	1	2	2	5	5	14	4	15	7	436	1	6	0	498
	0	0.01	0.03	0.03	0.07	0.07	0.21	0.06	0.22	0.1	6.41	0	0.09	0	7.32
2009	0	0	2	3	5	1	17	9	14	0	352	1	8	0	412
	0	0	0.03	0.04	0.07	0.01	0.25	0.13	0.21	0	5.17	0	0.12	0	6.05
2010	0	0	3	9	1	4	19	4	19	0	481	0	5	0	545
	0	0	0.04	0.13	0.01	0.06	0.28	0.06	0.28	0	7.07	0	0.07	0	8.01
2011	0	0	3	5	4	3	26	4	20	0	394	0	4	0	463
	0	0	0.04	0.07	0.06	0.04	0.38	0.06	0.29	0	5.79	0	0.06	0	6.8
Total	11	7	29	24	167	53	268	52	280	99	5718	2	93	2	6805
	0.16	0.1	0.43	0.35	2.45	0.78	3.94	0.76	4.11	1.45	84	0	1.37	0	100

www.manaraa.com

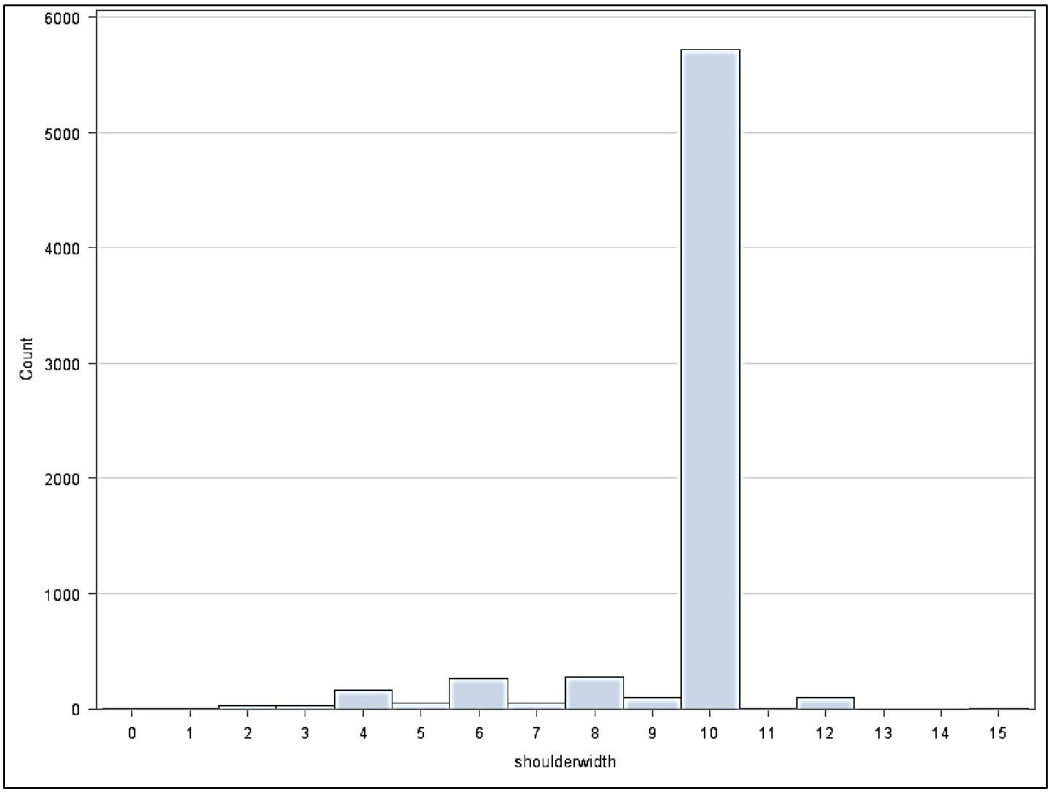

Figure 1. Histogram of the total number of segment observations for each class of shoulder width

www.manaraa.com

Table 2. Frequency and percentages for the newly defined shoulderwidth group within each year

year	shoulderwidthclass			
	1	2	3	Total
2002	82	983	11	1076
	1.2	14.5	0.16	15.8
2003	104	955	14	1073
	1.53	14	0.21	15.8
2004	123	936	18	1077
	1.81	13.8	0.26	15.8
2005	106	831	18	955
	1.56	12.2	0.26	14
2006	14	227	4	245
	0.21	3.34	0.06	3.6
2007	27	427	7	461
	0.4	6.27	0.1	6.77
2008	33	459	6	498
	0.48	6.75	0.09	7.32
2009	37	367	8	412
	0.54	5.39	0.12	6.05
2010	40	500	5	545
	0.59	7.35	0.07	8.01
2011	45	414	4	463
	0.66	6.08	0.06	6.8
Total	611	6099	95	6805
	8.98	89.6	1.4	100

www.manaraa.com

Table 3. Quantiles of the shoulderwidth variable

Quantiles (Shoulderwidth)	
Quantile	Estimate
100% Max	15
99%	12
95%	10
90%	10
75% Q3	10
50% Median	10
25% Q1	10
10%	8
5%	6
1%	3
0% Min	0

Examining PSR

PSR (pavement serviceability rate) which is a factor related to the pavement condition is a continuous variable and the only way to address the reviewers' comment was to categorize the observations according to its mean vale (32.474004) and standard deviation (2.6395756). In order to be able to obtain an optimal categorization across the maximum number of years, three classes of PSR was defined in the following format:

- Class 1, "PSR < (mean – 0.3 x standard deviation)",
- Class 2, "(mean – 0.3 x standard deviation) < PSR < (mean + 0.3 x standard deviation)", and
- Class 3, "(mean + 0.3 x standard deviation) < PSR".

Table 4 shows the number of observation in each category. It can be observed that three of the PSR categories by year groups lack enough data (based on the recommended 60 observations). This categorization was nevertheless used in the model as these anomalies occurred in a minor portion of the whole data set and the shortage in the number of observations is relatively minor.

www.manaraa.com

Table 4. Frequency in each newly defined PSR (pavement serviceability rate) categories

year	PSR class			
	1 (low)	2 (med)	3 (high)	Total
2002	445	187	420	1052
	6.76	2.84	6.38	15.98
2003	396	200	455	1051
	6.02	3.04	6.91	15.97
2004	439	196	427	1062
	6.67	2.98	6.49	16.13
2005	239	205	496	940
	3.63	3.11	7.53	14.28
2006	46	43	151	240
	0.7	0.65	2.29	3.65
2007	65	56	334	455
	0.99	0.85	5.07	6.91
2008	110	85	279	474
	1.67	1.29	4.24	7.2
2009	93	80	226	399
	1.41	1.22	3.43	6.06
2010	321	92	82	495
	4.88	1.4	1.25	7.52
2011	284	68	63	415
	4.31	1.03	0.96	6.3
Total	2438	1212	2933	6583
	37	18.4	44.6	100

Examining Number of Lanes (Nolanes)

First the number of observations for each number of lanes category were tabulated to see if there is a sufficient number of observations in each group. Table 5 shows the results of this tabularization. In each group there are two numbers presented in the table: namely the actual number of observations and the overall percentage for that group. For example there are 38 segments observed with nolanes=5 during the year 2002 which is 0.6% of the overall number of observations. It can be observed that the nolanes categories

www.manaraa.com

equal to 4, 5, 6, and 7 lanes lack enough observations (as suggested by the reviewer to be at least 60). For this reason it was decided to combine all these groups into one category of nolanes>3. Therefore, three groups of nolanes were determined as follows: those with nolanes=2, nolanes=3, and nolanes>3. Table 6 presents the new sets of categories and the number of observations (nolaneclass) in each one of them which shows that there are acceptable values as the sufficient number of observations.

Now there is another issue which is the confounding effect of the nolanes with the urbanrural variable. To illustrate this, another table that shows the distribution of the number of observations for each type of urban or rural is presented (Table 7).

Table 5. Frequency and percentages for each class of nolanes within each year

year	nolanes						
	2	3	4	5	6	7	Total
2002	672	212	140	38	12	2	1076
	9.88	3.12	2.06	0.6	0.2	0	15.81
2003	663	221	141	34	12	2	1073
	9.74	3.25	2.07	0.5	0.2	0	15.77
2004	658	220	147	46	6	0	1077
	9.67	3.23	2.16	0.7	0.1	0	15.83
2005	555	216	139	39	6	0	955
	8.16	3.17	2.04	0.6	0.1	0	14.03
2006	91	87	54	13	0	0	245
	1.34	1.28	0.79	0.2	0	0	3.6
2007	278	104	63	15	1	0	461
	4.09	1.53	0.93	0.2	0	0	6.77
2008	326	112	46	13	1	0	498
	4.79	1.65	0.68	0.2	0	0	7.32
2009	243	102	44	21	2	0	412
	3.57	1.5	0.65	0.3	0	0	6.05
2010	337	133	42	33	0	0	545
	4.95	1.95	0.62	0.5	0	0	8.01
2011	281	106	58	17	1	0	463
	4.13	1.56	0.85	0.3	0	0	6.8
Total	4104	1513	874	269	41	4	6805
	60.3	22.2	12.8	4	0.6	0.1	100

www.manaraa.com

Table 6. Observation frequency and percentages for the newly defined nolanes group within each year

year	nolaneclass			
	2	3	>3	Total
2002	672	212	192	1076
	9.88	3.12	2.82	15.81
2003	663	221	189	1073
	9.74	3.25	2.78	15.77
2004	658	220	199	1077
	9.67	3.23	2.92	15.83
2005	555	216	184	955
	8.16	3.17	2.7	14.03
2006	91	87	67	245
	1.34	1.28	0.98	3.6
2007	278	104	79	461
	4.09	1.53	1.16	6.77
2008	326	112	60	498
	4.79	1.65	0.88	7.32
2009	243	102	67	412
	3.57	1.5	0.98	6.05
2010	337	133	75	545
	4.95	1.95	1.1	8.01
2011	281	106	76	463
	4.13	1.56	1.12	6.8
Total	4104	1513	1188	6805
	60.31	22.23	17.46	100

www.manaraa.com

Table 7. Distribution of number of observations for each of the nolanes within the area types

Area type	nolanes						
	2	3	4	5	6	7	Total
Rural	2463	39	0	0	0	0	2502
	36.19	0.57	0	0	0	0	36.77
Urban	1641	1474	874	269	41	4	4303
	24.11	21.66	12.84	3.95	0.6	0.06	63.23
Total	4104	1513	874	269	41	4	6805
	60.31	22.23	12.84	3.95	0.6	0.06	100

It can be seen that the rural area includes only highways with 2 or 3 lanes. Moreover, there are only 39 observations in the nolanes=3 category for rural and also nolanes 6 and 7 for the urban have much fewer than 60 observations as recommended. Since the nolanes factor was converted into the grouping nolanes=2, =3 and >3, a table giving the sample size for each of these categories was obtained (Table 8).

Table 8. Distribution of number of observations for the newly defined nolanes group within area types

Area type	nolaneclass			
	2	3	>3	Total
Rural	2463	39	0	2502
	36.2	0.57	0	36.8
Urban	1641	1474	1188	4303
	24.1	21.7	17.5	63.2
Total	4104	1513	1188	6805
	60.3	22.2	17.5	100

The second reviewer suggested considering the urban/rural designation of the road segments in the model. Accommodating this as well as including a categorical variable for the number of lanes pose a challenge because of the confounding observed between the two variables. In other words, an effect associated with the number of lanes categorical variable cannot be separated from the effect due to the urban/rural variable.

www.manaraa.com

To overcome this problem, four categories were defined (under the variable name arealane): urban two lanes (arealane=1), urban three lanes (arealane=2), urban three plus lanes (arealane=3), and rural (arealane=4). The 39 rural road segments with 3 lanes were combined into rural segments with 2 lanes rather than delete these observations. One might say it should not be done, but a separate analysis was also conducted where that rural segments with 3 lanes were deleted from the dataset. No significant change was observed in the estimates (Tables 9 and 10).

Table 9. Model estimates with road segments with 2 and 3 lanes combined for rural type

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-11.548	0.8813	-13.2752	-9.8208	-13.1	<.0001
lanewidth		0.0271	0.0541	-0.079	0.1332	0.5	0.6168
shoulderwidth		-0.0005	0.0099	-0.0199	0.0189	-0.05	0.9596
arealane	1	1.2733	0.0934	1.0903	1.4563	13.6	<.0001
arealane	2	1.1949	0.108	0.9833	1.4065	11.1	<.0001
arealane	3	1.144	0.1178	0.9131	1.3749	9.71	<.0001
arealane	4	0	0	0	0	.	.
lnAADT		1.1702	0.0557	1.0609	1.2794	21	<.0001
SL		-0.0067	0.0017	-0.0102	-0.0033	-3.86	0.0001
PSR		-0.0045	0.005	-0.0142	0.0053	-0.9	0.3687
percentcommercial		-0.4511	0.2491	-0.9394	0.0371	-1.81	0.0702

Table 10. Model estimates with rural road segments with 3 lanes removed from dataset

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-11.4554	0.8793	-13.1789	-9.7319	-13.03	<.0001
lanewidth		0.028	0.0544	-0.0785	0.1346	0.52	0.6059
shoulderwidth		-0.0039	0.0094	-0.0224	0.0146	-0.42	0.6775
arealane	1	1.2946	0.0944	1.1096	1.4796	13.71	<.0001
arealane	2	1.2338	0.1068	1.0245	1.443	11.56	<.0001
arealane	3	1.1785	0.1174	0.9483	1.4087	10.03	<.0001
arealane	4	0	0	0	0	.	.
lnAADT		1.1603	0.0554	1.0518	1.2689	20.94	<.0001
SL		-0.0066	0.0017	-0.01	-0.0032	-3.79	0.0001
PSR		-0.0048	0.005	-0.0145	0.0049	-0.97	0.3329
percentcommercial		-0.4264	0.2508	-0.918	0.0651	-1.7	0.0891

www.manaraa.com

Examining Speed limit

Similar to the previous variables, speed limit was examined for the number of observations in each class per year (Table 11). It can be observed that a few of the groupings lack enough observations for some years.

Table 11. Frequency and percentages for each class of speed limit within each year

year	SL				
	55	60	65	70	Total
2002	113	223	137	579	1052
	1.72	3.39	2.08	8.8	15.98
2003	113	221	137	580	1051
	1.72	3.36	2.08	8.81	15.97
2004	140	218	123	581	1062
	2.13	3.31	1.87	8.83	16.13
2005	128	204	118	490	940
	1.94	3.1	1.79	7.44	14.28
2006	66	81	22	71	240
	1	1.23	0.33	1.08	3.65
2007	78	99	48	230	455
	1.18	1.5	0.73	3.49	6.91
2008	69	53	82	270	474
	1.05	0.81	1.25	4.1	7.2
2009	72	41	85	201	399
	1.09	0.62	1.29	3.05	6.06
2010	76	41	90	288	495
	1.15	0.62	1.37	4.37	7.52
2011	60	41	85	229	415
	0.91	0.62	1.29	3.48	6.3
Total	915	1222	927	3519	6583
	13.9	18.56	14.08	53.46	100

Similar to the nolanes variable, a tabularization was conducted for the SL classes versus the area type to see if there are similar patterns. It was again observed that the rural area lacks a sufficient number of observations in the SL classes of SL=55, 60, and 65 and 98.32% of the rural segments observed have a speed limit of 70 mph. This issue was not observed in the urban category.

www.manaraa.com

Table 12. Distribution of number of observations for each of the speed limit classes within the area types

Area type	Speed Limit				
	55	60	65	70	Total
Rural	0	28	14	2459	2501
	0	0.43	0.21	37.35	37.99
Urban	915	1194	913	1060	4082
	13.9	18.14	13.87	16.1	62.01
Total	915	1222	927	3519	6583
	13.9	18.56	14.08	53.46	100

One can say that there is confounding amongst the variables SL, nolanes, and the type of the area (urban or rural). Several different analyses were conducted using these newly defined categories to investigate the effect of each one of the categories (area, number of lanes, and speed limit) when fitted simultaneously as categorical variables. Since there is confounding, some effects and their interactions were not estimable. Such estimability issues arising out of confounding are to be expected. As a solution it was decided to define new dummy variables each of which represents one of the area*nolanes*SL combinations. Three groupings were chosen for the three nolanes categories of 2, 3, and 3+ lanes, Four SL categories of 55, 60, 65, and 70 mph with the exception that there was no observation for the rural area with 55 mph speed limit. The new categories (dummy variables) that were used in the model and the number of observation in each variable are presented in Table 13. In the name of the dummy variables, first part indicates the area type criterion, second part indicates the number of lanes criterion and the third part indicates the speed limit criterion.

www.manaraa.com

Table 13 Name and number of observations of the Combinatory dummy variables

Category	Number of observation
urban_2_55	699
urban_2_60	180
urban_2_65	261
urban_2_70	877
urban_3_55	415
urban_3_60	323
urban_3_65	484
urban_3_70	140
urban_3p_55	216
urban_3p_60	691
urban_3p_65	168
urban_3p_70	43
rural_2_60	28
rural_2_65	12
rural_2_70	2422
rural_3_60	0
rural_3_65	2
rural_3_70	37

So there are overall 12 categories defined for urban and 6 categories for the rural area segments. The group in rural area with 3 lanes and speed limit of 60 mph (rural_3_60) had zero observation and therefore not used in the model. The rural category with 2 lanes and 70 mph (rural_2_70) was used as the base condition in the model. It should also be mentioned that the other categories in the rural arena did not have our target value of 60 observations but they were retained to avoid removing the data. The soundness of this decision was double checked by running two models, one with and another without the small-sized rural variables and comparing the two model estimates (Table and Table).

From the results it is seen that all the dummy variables in the rural category were not found to be significant variables in the model. In other words, these categories are not

www.manaraa.com

resulting in statistically different effects from the base condition which is the rural_2_70. This might be because of the small number of observations that exist in those categories. Therefore, all those rural categories were lumped together with the rural_2_70 group.

This result also indicates that there is an overall significant difference between the urban and rural areas and how they affect the crash occurrences (higher frequency in urban areas). A similar pattern was also observed for the tradition NB model. Table 15 show the refined results for the GEE model wherein, only the urban categories were considered in the model.

Table 14. Analysis of parameter estimates using generalized estimating equations with the rural variables in the model

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-11.7718	0.9555	-13.6445	-9.8991	-12.32	<.0001
lnAADT	1.1278	0.0642	1.0019	1.2537	17.56	<.0001
Psrclass 1 (low)	-0.0079	0.0273	-0.0614	0.0456	-0.29	0.7725
Psrclass 2 (med)	-0.03	0.0272	-0.0834	0.0234	-1.1	0.2712
Psrclass 3 (high)	0	0	0	0	.	.
Percentcommercial	-0.3222	0.2601	-0.832	0.1876	-1.24	0.2154
Lanewidth	0.0237	0.0592	-0.0923	0.1397	0.4	0.689
urban_2_55	1.7909	0.1327	1.5308	2.051	13.49	<.0001
urban_2_60	1.3385	0.1659	1.0134	1.6637	8.07	<.0001
urban_2_65	1.348	0.1177	1.1172	1.5787	11.45	<.0001
urban_2_70	1.1648	0.1035	0.9618	1.3677	11.25	<.0001
urban_3_55	1.5297	0.1361	1.2628	1.7965	11.24	<.0001
urban_3_60	1.4034	0.1489	1.1116	1.6952	9.43	<.0001
urban_3_65	1.2075	0.1232	0.9659	1.449	9.8	<.0001
urban_3_70	1.0119	0.1922	0.6353	1.3886	5.27	<.0001
urban_3p_55	1.504	0.1586	1.1932	1.8148	9.49	<.0001
urban_3p_60	1.2735	0.1381	1.0028	1.5442	9.22	<.0001
urban_3p_65	1.323	0.1381	1.0523	1.5938	9.58	<.0001
urban_3p_70	0.9379	0.2723	0.4041	1.4717	3.44	0.0006
rural_2_60	0.3894	0.3343	-0.2658	1.0447	1.16	0.2441
rural_2_65	0.1594	0.2113	-0.2548	0.5736	0.75	0.4507
rural_3_65	0.5008	0.7014	-0.874	1.8755	0.71	0.4753
rural_3_70	0.0761	0.2965	-0.505	0.6571	0.26	0.7975

www.manaraa.com

Table 15. Analysis of parameter estimates using generalized estimating equations without the rural variables in the model

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-11.7783	0.9549	-13.6499	-9.9068	-12.33	<.0001
lnAADT	1.1295	0.0642	1.0037	1.2553	17.6	<.0001
psrclass	-0.0106	0.0273	-0.0641	0.0429	-0.39	0.6983
psrclass	-0.0322	0.0274	-0.086	0.0216	-1.17	0.2407
psrclass	0	0	0	0	.	.
percentcommercial	-0.3285	0.2602	-0.8384	0.1815	-1.26	0.2068
lanewidth	0.0239	0.0591	-0.092	0.1397	0.4	0.6865
urban_2_55	1.7816	0.1322	1.5225	2.0408	13.47	<.0001
urban_2_60	1.3297	0.1656	1.0052	1.6542	8.03	<.0001
urban_2_65	1.338	0.1173	1.108	1.5679	11.4	<.0001
urban_2_70	1.1583	0.1024	0.9577	1.3589	11.32	<.0001
urban_3_55	1.519	0.1355	1.2535	1.7846	11.21	<.0001
urban_3_60	1.3922	0.1483	1.1015	1.6828	9.39	<.0001
urban_3_65	1.1952	0.1239	0.9524	1.438	9.65	<.0001
urban_3_70	1.0006	0.1915	0.6253	1.3758	5.23	<.0001
urban_3p_55	1.4934	0.158	1.1836	1.8031	9.45	<.0001
urban_3p_60	1.2618	0.1373	0.9926	1.5309	9.19	<.0001
urban_3p_65	1.3114	0.1377	1.0414	1.5814	9.52	<.0001
urban_3p_70	0.9277	0.272	0.3945	1.4609	3.41	0.0006

In order to allow the other main variables in the model to have different impact on accident count in urban and rural road segments, a new dummy variable "area" set to be 0 for rural (base) and 1 being urban and the interactions of this variable with the other main factors of the model were also included in the model:

Areadt = area x lnAADT;

Areacommercial = area x percentcommercial;

Areawidth = area x lanewidth;

Areapsr = area x psr;

This will allow us to determine if the main factors influence crash frequency differentially across area. From these new terms, the areawidth resulted in a complicated

www.manaraa.com

convergence iteration process that did not satisfy the convergence criterion. Investigating the number of observations per lanewidth category for the rural and urban areas revealed that only 3 of 7 lane width levels were observed in the rural area and a complete examination of the interaction effect was not possible with the available data. Therefore, this term was removed from the model. The interaction term Areapsr was not found to be statistically significant in any of the two models, GEE and MLE models (traditional NB model) and was also removed from the analysis; however, the main factors that were not found statistically significant were left to remain in the model. The final model included the following variables:

- LnAADT
- psrclass
- percentcommercial
- lanewidth
- areadt
- areacommercial
- urban_2_55, urban_2_60, urban_2_65, urban_2_70
- urban_3_55, urban_3_60, urban_3_65, urban_3_70
- urban_3p_55, urban_3p_60, urban_3p_65, urban_3p_70

www.manaraa.com

APPENDIX D.

PAPER CORRECTIONS ADDENDUM

www.manaraa.com

This section was added to the dissertation to address the comments received from the committee members during the defense session. Since this dissertation is paper-based and at the time of defense the first and second papers are already published and in-press, respectively, the comments are addressed as an addendum to the dissertation file.

Comments for paper I

- Abstract section, line 1, changes to "This study systematically evaluates the changes in the frequency of motor vehicle crashes that…"
- Abstract section, line 1, changes to "…following the implementation of Missouri's *Strategic Highway Safety Plan (MSHSP) between 2005 and 2007"*
- Abstract section, line 16, changes to "The empirical results indicate that the MSHSP was a successful…"
- Introduction section, paragraph 2, line 9, changes to "The present study empirically examines the effect of implementation of the…"
- Introduction section, paragraph 1, line 5, changes to "The potentially life-saving and injury-reducing strategies in Missouri's Blueprint…"
- Background section, paragraph 1, line 1, changes to "Highway safety analysts use regression models for purposes such as estimating relationships between motor…"
- Methodology section, last paragraph, the following bibliography should be cited at the end of the first sentence related to the phenomenon of regression to the mean in negative binomial models:

  Maher, M. 1990. A bivariate negative binomial model to explain traffic accident migration. Accident Analysis and Prevention, Vol. 22(5), 487-498.

- Caption of Table 3 changes to "Comparison of the predicted crash count per segment properties for 2008 with/without safety improvements"
- The first heading inside the Table 3 changes from "Models for all collision types combined" to "models for different severity levels combined over all collision types"
- The second heading inside the Table 3 changes from "Models for all severity levels combined" to "models for different collision types combined over all severity levels"

Comments for paper II

- Caption of Figure 1 changes to "Total number of crashes on a select few of the interstate highways of Missouri with most variation. (legend presents the name of the interstate highways, e.g. 44 indicates interstate 44)"

www.manaraa.com

- Caption of Figure 5 changes to "Cumulative residuals plot for LnAADT for the negative binomial models estimated using generalized estimating equation methods (top) and maximum likelihood estimation (bottom)"
- Caption of Table 1 changes to "Correlations for the autoregressive Type 1 and exchangeable structure"
- Abstract section, sentence 1, changes to "The prediction in crash frequency models can be…"
- Abstract section, sentence 1, changes to "… years of data most commonly used in the literature"
- Abstract section, sentence 4, changes to "Despite the obvious temporal correlation of crashes, analyses of such correlation have been limited and the consequences of omitting temporal correlation are not completely understood"
- Introduction section, paragraph 2, sentence 7, changes to "…omission of the serial correlation are still not completely understood"
- Introduction section, paragraph 3, sentence 3, changes to "The GEE approach treats each highway segment as a cluster…"
- Introduction section, paragraph 4, sentence 1, changes to "Lord and Mahlawat (2009) used a GEE method with an…"

www.manaraa.com

BIBLIOGRAPHY

Aarts, L., and van Schagen, I. 2006. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention,* Vol. *38*(2), 215-224. doi: http://dx.doi.org/10.1016/j.aap.2005.07.004

Abdel-Aty, M., and Radwan, A. E. 2000. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention,* Vol. *32*(5), 633-642.

Aguero-Valverde, J., and Jovanis, P. P. 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. *2136*, 82-91.

Ahmed, M., Huang, H., Abdel-Aty, M., and Guevara, B. 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis and Prevention,* Vol. *43*(4), 1581-1589.

Allison, P. D. 2012. *Logistic regression using SAS: Theory and application*: SAS Institute.

Anastasopoulos, P. C., and Mannering, F. L. 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. *Accident Analysis & Prevention,* Vol. *43*(3), 1140-1147.

ASTM-D6433-07. 2007. Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys. West Conshohocken, PA: ASTM International.

ASTM-E1489-08. 2008. Standard Practice for Computing Ride Number of Roads from Longitudinal Profile Measurements Made by an Inertial Profile Measuring Device. West Conshohocken, PA: ASTM International.

Ballinger, G. A. 2004. Using generalized estimating equations for longitudinal data analysis. *Organizational research methods,* Vol. *7*(2), 127-150.

Barnett, A. G., van der Pols, J. C., and Dobson, A. J. 2005. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology,* Vol. *34*(1), 215-220.

Belsley, D. A., Kuh, E., and Welsch, R. E. 2005. *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571): John Wiley and Sons.

Bhat, C. R., Born, K., Sidharthan, R., and Bhat, P. C. 2014. A count data model with endogenous covariates: Formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research,* Vol. *1*, 53-71.

Buddhavarapu, P., Banerjee, A., and Prozzi, J. A. 2013. Influence of pavement condition on horizontal curve safety. *Accident Analysis & Prevention,* Vol. *52*, 9-18.

Carson, J., and Mannering, F. L. 2001. The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis and Prevention,* Vol. *33*(1), 99-109.

www.manaraa.com

Castro, M., Paleti, R., and Bhat, C. R. 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B,* Vol. *46*(1), 253-272.

CERS. (2010). National Rural Road Safety Public Opinion Survey Accessed July 24, 2014, from http://www.ruralsafety.umn.edu/publications/nationalsafetysurvey/index.html

Chang, H. L., Woo, T. H., and Tseng, C. M. 2006. Is rigorous punishment effective? A case study of lifetime license revocation in Taiwan. *Accident Analysis and Prevention,* Vol. *38*(2), 269-276.

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

Chi, G., McClure, T. E., and Brown, D. B. 2012. Gasoline prices and traffic crashes in Alabama, 1999–2009. *Traffic Injury Prevention,* Vol. *13*(5), 476-484.

Cox, D. R. 1984. Interaction. *International Statistical Review/Revue Internationale de Statistique,* Vol., 1-24.

Datta, T. K., Schattler, K., and Datta, S. 2000. Red light violations and crashes at urban intersections. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. *1734*(1), 52-58.

Dong, C., Clarke, D. B., Richards, S. H., and Huang, B. 2014. Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: An exploratory analysis. *Accident Analysis and Prevention,* Vol. *62*, 87-94.

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

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

Dupont, E., Papadimitriou, E., Martensen, H., and Yannis, G. 2013. Multilevel analysis in road safety research. *Accident Analysis and Prevention,* Vol.

Elvik, R. 2000. How much do road accidents cost the national economy? *Accident Analysis & Prevention,* Vol. *32*(6), 849-851.

Elvik, R., Christensen, P., and Amundsen, A. 2004. Speed and road accidents. *An evaluation of the Power Model. TØI report,* Vol. *740*, 2004.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. 1993. Regression models for discrete longitudinal responses. *Statistical Science,* Vol., 284-299.

Garber, N., and Hoel, L. 2008a. *Traffic & highway engineering*: Cengage Learning.

Garber, N., and Hoel, L. 2008b. *Traffic and highway engineering*: Cengage Learning.

www.manaraa.com

Gill, J. 2001. *Generalized linear models: a unified approach* (Vol. 134): Sage Publications, Incorporated.

Giuffrè, O., Granà, A., Giuffrè, T., and Marino, R. 2007. Improving reliability of road safety estimates based on high correlated accident counts *Transportation Research Record: Journal of the Transportation Research Board* (Vol. 2019, pp. 197-204).

Giuffrè, O., Grana, A., Giuffrè, T., and Marino, R. 2013. Accounting for Dispersion and Correlation in Estimating Safety Performance Functions. An Overview Starting from a Case Study *Modern Applied Science* (Vol. 7, pp. p11).

Guo, F., Wang, X., and Abdel-Aty, M. 2010a. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis and Prevention,* Vol. *42*(1), 84-92.

Guo, F., Wang, X., and Abdel-Aty, M. A. 2010b. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention,* Vol. *42*(1), 84-92.

Hanley, J. A., Negassa, A., and Forrester, J. E. 2003. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology,* Vol. *157*(4), 364-375.

Hardin, J. W., and Hilbe, J. M. (2007). Generalized Estimating Equations *Wiley Encyclopedia of Clinical Trials*: John Wiley and Sons, Inc.

Hauer, E., and Persaud, B. 1987. *How to estimate the safety of rail-highway grade crossings and the safety effects of warning devices*: Transportation Research Board.

Hauer, E. 1997. *Observational Before/After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*.

Hauer, E., and Bamfo, J. 1997. *Two tools for finding what function links the dependent variable to the explanatory variables.* Paper presented at the Proceedings of the ICTCT 1997 Conference.

Hilton, B. N., Horan, T. A., Burkhard, R., and Schooley, B. 2011. SafeRoadMaps: Communication of location and density of traffic fatalities through spatial visualization and heat map analysis. *Information Visualization,* Vol. *10*(1), 82-96.

HSM. 2010. *Highway Safety Manual* (Vol. 1): AASHTO.

Hutchings, C. B., Knight, S., and Reading, J. C. 2003. The use of generalized estimating equations in the analysis of motor vehicle crash data. *Accident Analysis and Prevention,* Vol. *35*(1), 3-8.

Jung, S., Xiao, Q., and Yoon, Y. 2013a. Evaluation of motorcycle safety strategies using the severity of injuries. *Accident Analysis & Prevention,* Vol. *59*, 357-364.

Jung, S., Xiao, Q., and Yoon, Y. 2013b. Evaluation of motorcycle safety strategies using the severity of injuries. *Accident Analysis and Prevention,* Vol. *59*, 357-364.

www.manaraa.com

Kempton, W., Brown, M., Murphy, C. J., Valverde, G., and Jolly, J. R. 2006. California Strategic Highway Safety Plan, Version 2. California Business, Transportation, Housing Agency Contributing Departments, Sacramento, CA. .

Khorashadi, A., Niemeier, D., Shankar, V., and Mannering, F. 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. *Accident Analysis & Prevention,* Vol. *37*(5), 910-921.

Knuiman, M. W., Council, F. M., and Reinfurt, D. W. 1993. Association of median width and highway accident rates. *Transportation Research Record,* Vol., 70-70.

Kutner, M. H., Nachtsheim, C., and Neter, J. 2004. Applied linear regression models. Vol.

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

Lenguerrand, E., Martin, J. L., and Laumon, B. 2006. Modelling the hierarchical structure of road crash data—Application to severity analysis. *Accident Analysis and Prevention,* Vol. *38*(1), 43-53.

Li, X., Lord, D., Zhang, Y., and Xie, Y. 2008. Predicting motor vehicle crashes using support vector machine models. *Accident Analysis and Prevention,* Vol. *40*(4), 1611-1618.

Liang, K.-Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika,* Vol. *73*(1), 13-22.

Lin, D., Wei, L., and Ying, Z. 2002. Model-Checking Techniques Based on Cumulative Residuals. *Biometrics,* Vol. *58*(1), 1-12.

Littell, R. C., Stroup, W. W., and Freund, R. J. 2002. *SAS for linear models*: SAS Institute.

Lord, D., and Persaud, B. N. (2000). Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record: Journal of the Transportation Research Board, 1717*, 102-108. Retrieved from

Lord, D., and Park, P. Y.-J. 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis and Prevention,* Vol. *40*(4), 1441-1457. doi: http://dx.doi.org/10.1016/j.aap.2008.03.014

Lord, D., and Mahlawat, M. 2009. Examining Application of Aggregated and Disaggregated Poisson-Gamma Models Subjected to Low Sample Mean Bias. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. *2136*, 1-10.

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

www.manaraa.com

Maas, C. J., and Hox, J. J. 1999. Sample sizes for multilevel modeling. *Am J Public Health,* Vol. *89*, 1181-1186.

Maher, M. J., and Summersgill, I. 1996. A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis and Prevention,* Vol. *28*(3), 281-296.

Mancl, L. A., and DeRouen, T. A. 2001. A covariance estimator for GEE with improved small-sample properties. *Biometrics,* Vol. *57*(1), 126-134.

Mannering, F. L., and Bhat, C. R. 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research,* Vol. *1*, 1-22.

Manuel, A., El-Basyouny, K., and Islam, M. T. 2014. Investigating the safety effects of road width on urban collector roadways. *Safety science,* Vol. *62*, 305-311.

Martens, M., Compte, S., and Kaptein, N. A. 1997. The effects of road design on speed behaviour: a literature review. Vol.

McCullagh, P., and Nelder, J. A. 1989. *Generalized linear model* (Vol. 37): Chapman and Hall/CRC.

Méndez, Á. G., Aparicio Izquierdo, F., and Ramírez, B. A. 2010. Evolution of the crashworthiness and aggressivity of the Spanish car fleet. *Accident Analysis and Prevention,* Vol. *42*(6), 1621-1631.

Milton, J. C., and Mannering, F. L. 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation,* Vol. *25*(4), 395-413.

MoDOT. (2004). Missouri's blueprint for safer roadways Accessed July 26, 2014, from http://www.ite.org/safety/stateprograms/Missouri_SHSP.pdf

MoDOT. (2008). Missouri's blueprint to arrive alive Accessed July 26, 2014, from http://www.savemolives.com/documents/FINALBlueprintdocument.pdf

Mohammadi, M. A., Samaranayake, V. A., and Bham, G. 2013. *The Effect of Incorporating Temporal Correlations into Negative Binomial Count Data Models.* Paper presented at the Fourth International Conference on Road Safety and Simulation, Rome, Italy.

Mohammadi, M. A. (2014). *Longitudinal analysis of crash frequency data.* Doctoral dissertation, Missouri University of Sicence and Technology.

Mohammadi, M. A., Samaranayake, V., and Bham, G. H. 2014. *Safety Effect of Missouri's Strategic Highway Safety Plan-Missouri's Blueprint for Safer Roadways.* Paper presented at the Transportation Research Board 93rd Annual Meeting.

Mohammadi, M. A., Samaranayake, V. A., and Bham, G. 2014a. *Safety Effect of Missouri's Strategic Highway Safety Plan-Missouri's Blueprint for Safer Roadways.* Paper presented at the Transportation Research Board 93rd Annual Meeting.

www.manaraa.com

Mohammadi, M. A., Samaranayake, V. A., and Bham, G. 2014b. Crash Frequency Modeling using Negative Binomial Models: An Application of Generalized Estimating Equation to Longitudinal Data. *Accepted for publication in Analytic Methods in Accident Research,* Vol. *2*.

Nelder, J. A. 1977. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General),* Vol., 48-77.

NHTSA. (2008). *Traffic safety facts*. National Highway Transportation Safety Association.

NHTSA. (2009). *Traffic safety facts*. Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/811402.pdf.

Noland, R. B., and Oh, L. 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. *Accident Analysis and Prevention,* Vol. *36*(4), 525-532.

Noland, R. B., Quddus, M. A., and Ochieng, W. Y. 2008. The effect of the London congestion charge on road casualties: an intervention analysis. *Transportation,* Vol. *35*(1), 73-91.

Pan, W. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics,* Vol. *57*(1), 120-125.

Park, B.-J., and Lord, D. 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention,* Vol. *41*(4), 683-691. doi: http://dx.doi.org/10.1016/j.aap.2009.03.007

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., and Mathers, C. D. 2004. World report on road traffic injury prevention: World Health Organization Geneva.

Peng, Y., Boyle, L. N., and Hallmark, S. L. 2012. Driver's lane keeping ability with eyes off road: Insights from a naturalistic study. *Accident Analysis and Prevention,* Vol.

Persaud, B., Retting, R. A., and Lyon, C. A. 2004. Crash reduction following installation of centerline rumble strips on rural two-lane roads. *Accident Analysis and Prevention,* Vol. *36*(6), 1073-1079.

Persaud, B., and Lyon, C. 2007. Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. *Accident Analysis & Prevention,* Vol. *39*(3), 546-555.

Poch, M., and Mannering, F. L. 1996. Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering,* Vol. *122*(2), 105-113.

Quddus, M. A. 2008. Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention,* Vol. *40*(5), 1732-1741.

Roque, C., and Cardoso, J. L. 2014. Investigating the relationship between run-off-the-road crash frequency and traffic flow through different functional forms. *Accident Analysis and Prevention,* Vol. *63*, 121-132.

www.manaraa.com

SAS. 2008. *SAS/STAT 9.2 User's Guide: The GENMOD Procedure (book Excerpt)*: SAS Institute.

Savolainen, P. T., and Tarko, A. P. 2005. Safety impacts at intersections on curved segments. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. *1908*, 130-140.

Shankar, V. N., Mannering, F. L., and Barfield, W. 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention,* Vol. *27*(3), 371-389.

Shively, T. S., Kockelman, K., and Damien, P. 2010. A Bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. *Transportation research part B: methodological,* Vol. *44*(5), 699-715.

Squires, C. A., and Parsonson, P. S. 1989. Accident comparison of raised median and two-way left-turn lane median treatments. *Transportation Research Record,* Vol. *1239*, 30-40.

Stavrinos, D., Jones, J. L., Garner, A. A., Griffin, R., Franklin, C. A., Ball, D., Welburn, S. C., Ball, K. K., Sisiopiku, V. P., and Fine, P. R. 2013. Impact of distracted driving on safety and traffic flow. *Accident Analysis and Prevention,* Vol.

Ulfarsson, G. F., and Shankar, V. N. 2003. Accident count model based on multiyear cross-sectional roadway data with serial correlation. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. *1840*, 193-197.

Venkataraman, N., Ulfarsson, G. F., Shankar, V. N., Oh, J., and Park, M. 2011. Model of Relationship Between Interstate Crash Occurrence and Geometrics. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. *2236*, 41-48.

Venkataraman, N., Ulfarsson, G. F., and Shankar, V. N. 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis and Prevention,* Vol. *59*, 309-318.

Venkataraman, N., Shankar, V. N., Ulfarsson, G. F., and Deptuch, D. 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. *Analytic Methods in Accident Research,* Vol. *2*, 12-20.

Wang, X., and Abdel-Aty, M. 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accident Analysis and Prevention,* Vol. *38*(6), 1137-1150.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. 2011. *Statistical and econometric methods for transportation data analysis*: CRC press.

Xiong, Y., Tobias, J. L., and Mannering, F. L. 2014. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. *Transportation Research Part B,* Vol. *67*, 109-128.

www.manaraa.com

Yang, H., Ozbay, K., Ozturk, O., and Yildirimoglu, M. 2013. Modeling work zone crash frequency by quantifying measurement errors in work zone length. *Accident Analysis and Prevention,* Vol. *55*, 192-201.

Yu, R., Abdel-Aty, M., and Ahmed, M. 2013a. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis and Prevention,* Vol. *50*, 371-376.

Yu, R., Abdel-Aty, M., and Ahmed, M. 2013b. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention,* Vol. *50*, 371-376.

Zegeer, C. V., and Deacon, J. A. 1987. Effect of lane width, shoulder width, and shoulder type on highway safety. *State-of-the-Art Report,* Vol. (6).

Zegeer, C. V., Stewart, J. R., Huang, H. H., and Lagerwey, P. A. 2002. Safety effects of marked vs. unmarked crosswalks at uncontrolled locations: Executive summary and recommended guidelines.

Zeger, S. L., and Liang, K.-Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics,* Vol., 121-130.

Zhang, Y., Xie, Y., and Li, L. 2012. Crash frequency analysis of different types of urban roadway segments using generalized additive model. *Journal of Safety research,* Vol. *43*(2), 107-114.

Zorn, C. J. 2001. Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science,* Vol., 470-490.

Zou, Y., Zhang, Y., and Lord, D. 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. *Analytic Methods in Accident Research,* Vol. *1*, 39-52.

www.manaraa.com

VITA

Mojtaba Ale Mohammadi is currently a PhD student of Transportation Engineering at the Missouri University of Science and Technology (Missouri S&T), Rolla, Missouri. He was born in Ramsar, Iran. He Attended University of Tabriz from 2001 to 2005 and received his B.S. degree in Civil Engineering. In 2005 he entered the University of Tehran Polytechnic and earned his M.S. degree in Pavement and Transportation Engineering. He joined Missouri S&T in 2009 to pursue his PhD. During his PhD program he has worked on several projects such as traffic control and operations in work zones, intelligent transportation systems, and for his PhD dissertation on safety analysis and crash count modeling.

www.manaraa.com