A MACRO-LEVEL ANALYSIS OF SAFETY DATA USING GEOSPATIAL TECHNIQUES AND SPATIAL ECONOMETRIC METHODS AND MODELS

by

SAMWEL OYIER ZEPHANIAH

STEVEN L. JONES, COMMITTEE CHAIR
ALEXANDER HAINEN
RANDY SMITH
JAY LINDLY
JOE WEBER

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Civil, Construction and Environmental Engineering
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2017

ABSTRACT

Motor vehicle accidents are a source of many preventable injuries and deaths, worldwide. Several statistical and econometric models have been developed to predict and explain crash events. Research indicate that 93% of traffic accidents are due to human error. The objective of this research is twofold – first, to develop a macro level safety planning framework by identifying socioeconomic factors that influence crash frequencies and second, to characterize traffic congestion attributed to a crash events. To this effect, a Geographically Weighted Poisson Regression (GWPR) model, a suite of Spatial Econometric models and a Mixed Logit model were estimated. Data used included crash records from 2009 to 2013 in Alabama comprising 647,477 crash events. These included 4,814 crashes on Interstate 65 and 21,818 crashes related to Driving Under the Influence (DUI). Other data comprised socioeconomic data from US census, weather data, traffic data, spatial data from ESRI and crowd sourced speed data. Results indicate that DUI crash rates and frequencies at postal code level are predominantly influenced by rate of employment, income, population density, level of education, household size and housing characteristics. In addition, level of congestion attributed to a crash depends on factors including traffic volume, speed, weather, time of the event, severity of the crash, presence of physical barrier separating opposing traffic lanes, work zone, percent of heavy trucks and whether the crash occurred in an urban area or rural area. These results are unequivocal regarding the importance of geographic variation and heterogeneity in driver behavior and the general road safety.

DEDICATION

To my wife Beryl, and Sons: Jayson and Jerome

LIST OF ABBREVIATIONS AND SYMBOLS

AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
CARE	Critical Accident Reporting Environment
CAPS	Center for Advanced Public Safety
DUI	Driving Under Influence
IIA	Independence of irrelevant alternatives
LL	Log likelihood
ML	Maximum likelihood
MNL	Multinomial logit
WHO	World Health Organization
VMT	Vehicle Miles Travelled
GIS	Geographic Information Science
GLM	Generalized Linear Model
BLUE	Best Linear Unbiased Estimate
AADT	Annual Average Daily Traffic
SDMH	Speed Differential Mile Hour
TMC	Traffic Messaging Channel
SARMA	Spatial Autoregressive Moving Average
SDARMA	Spatial Durbin Autoregressive Moving Average
OLS	ordinary Least Square

CR	Crash Rate
SDM	Spatial Durbin Model
SDEM	Spatial Durbin Error Model
SMA	Spatial Moving Average
SDMA	Spatial Durbin Moving Average
SAC	Spatial Autoregressive Moving Average
SDAC	Spatial Durbin Autoregressive Moving Average
LISA	Local Indicator of Spatial Autocorrelation
ALDOT	Alabama Department of Transport
KDE	kernel Density Estimation
SAS	Statistical Analysis Software
GWR	Geographically Weighted Regression
GWPR	Geographically Weighted Poisson Regression
MSPE	Mean Square Prediction Error
MAD	Mean Absolute Deviance
MPB	Mean Prediction Bias
ANN	Average Nearest Neighbor
w_{ij}	Geographical weight for independent variable at location j with
d_{ij}	Euclidian distance between $i\ and\ j$.
$\theta_{i(k)}$	Adaptive bandwidth
(u_i, v_i)	geographic coordinates of region i.
y_i	Dependent variable in region i.
φ	Vector of parameters of chosen density

β	Coefficients to be determined
f	Probability density function
P	Probability
X	Explanatory variable

ACKNOWLEDGEMENTS

Special thanks to Dr. Steven L. Jones, my dissertation adviser and committee chair. For his leadership, proactive approach to research, deep and broad understanding of transportation and for his “human” nature. Dr. Jones helped me in many ways throughout my graduate studies at the University of Alabama. THANK YOU!

Dr. Hainen Alexander, thank you so much for your help with data preparation and econometric modeling particularly in Chapter 4. Dr. Joe Weber, thank you so much for your help with geospatial techniques and analysis, especially in Chapter 1 and 2. Dr. Holt, from University of Alabama business school, thank you for insightful assistance with econometric analysis. Dr. Jay Lindly, thank you for being in my dissertation committee and your great inspiring leadership. Dr. Randy Smith, thank you for being in my committee and facilitating availability of data for this research. Ms. Connie Harris, it was great to work with you. Thank you for your selfless support.

To my colleagues, Abhay Lidbe, Kofi Adanu, Gaurav Mehta, Irina Riehle, Md Abu Sufian Talukder, Naima Islam, Preston Jute, Elizabeth Connell, Tedla, Elsa, Luana Ozelim, Michael Dun, thank you for being in my research team and for a great intellectual engagement.

To my wife Beryl, thank you for everything. You know it. Your patience is golden. My sons Jayson and Jerome thank you for the awesome sacrifice. I know you missed me while I was in school.

Finally, I thank God for the blessing of life, health and family.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 General

Motor vehicle crashes are a source of many preventable deaths and injuries worldwide (World Health Organization, 2015). Even though the number of fatalities per 100 vehicle miles traveled (VMT) is declining in the United States, traffic crashes still result in over 30,000 deaths annually (National Highway Traffic Safety Administration, 2010). As a result, research in road safety has become an increasingly important topic, in which one of the challenges has been how to effectively analyze individual crash events. Several statistical models and econometric models have been developed to predict and explain crash events (Lord and Mannering, 2010). Besides exploring the physical infrastructure, research indicate that approximately 93% of crashes are due to human factors (Driggs-Campbell, et al., 2015). In addition, geographic heterogeneity in human attributes makes it more complex to identify the specific human factors that contribute to crash events. Geographic Information Science (GIS) has been used to visualize, analyze and interpret patterns in crash data. Notwithstanding, a continued effort in traffic analysis is required to identify high impact countermeasures. Whereas extensive research has been done on road safety regarding road geometry, weather condition and traffic volume, there are more research efforts seeking to quantify and establish the relationship between crash frequencies and socioeconomic factors. This has resulted in the need to incorporate safety planning in infrastructure planning to help identify intervention measures at planning stage.

Noting that motor vehicle crashes are random events which occur in space and time, they also exhibit an attribute of spatial dependence and spatial autocorrelation which should be considered in analysis. Most statistical and econometric models assume neither spatial correlation nor spatial dependence (e.g., Mehta and Lou, 2013; Islam, et al., 2014). Gauss-Markov assumptions require that there should be no multicollinearity, otherwise, the parameter estimates cannot be Best Linear Unbiased Estimate (BLUE).

Traditionally, all econometric models are anchored on a Generalized Linear Model (GLM) technique after variable transformation depending on the nature of the distribution and skewness of the data. Due to the magnitude of unobserved heterogeneity (Mannering, et al., 2016), most statistical and econometric models hardly comply with the homoscedasticity assumption which requires the residuals to be independently and irrelevantly distributed with a mean of zero and a given specific standard deviation. As a result, heteroskedasticity and serial autocorrelation cannot be rejected. Holding the assumption of neither spatial dependence nor serial autocorrelation validates the coefficient estimates. If these assumptions are lifted, then the estimates will always be biased. Due to this problem, spatial econometric methods and models offer an alternative methodology to analyze crash data. With the improvement in GIS technology and detailed data being collected for every crash event, new opportunities to analyze, understand and improve safety become available. Whereas, a detailed review and assessment of methodological alternatives has been done (Lord and Mannering, 2010), other studies have also shown that spatially informed analysis hold promise in developing a better understanding of road safety (Mannering and Bhat, 2014; Mehta, et al., 2014).

1.2 Overall Research Objective

There are two broad objectives of this research. The first main research objective is to identify and understand spatial relationships between macro level socioeconomic factors and crashes attributable to human behavior (e.g., driving under the influence). To achieve this objective, the research focuses on two separate but related, efforts. The first effort identifies macro level socioeconomic factors that significantly influence driving under the influence (DUI) crashes in Alabama using a Geographically Weighted Poisson Regression. The second effort examines the relationship between socioeconomic factors and traffic safety using a suite of spatial econometric models and as a case study on DUI crashes in Alabama. Both studies address the fact that there is spatial dependence among crash data. The findings from the first and the second sub-objectives are then compared to identify factors that can be used to improve the overall road safety.

The second main research objective is to examine the relationship between factors contributing to crash occurrence and severity to the extent of traffic congestion resulting from such events. To achieve this objective, the research focused on measurement and analysis of traffic congestion attributable to Interstate crashes and identify factors that influence the level of congestion.

1.3 Organization and Structure of the Dissertation

Chapter 1 presents an overall introduction and background of the application of geospatial techniques in road safety. Chapter 2 describes the development of a Geographically Weighted Poisson Regression (GWPR) model using aggregated postal code level crash frequency as dependent variable and socioeconomic data as independent variables. In this case, a geographic weighting is based on the average nearest neighborhood analysis which used optimum bandwidth and kernel function to estimate a postal code level GWPR model for crash frequency. Chapter 3 describes the application of spatial econometric methods and model to estimate various spatial models for characterizing crash rates at the postal code level. The geographic weighting in this case is based on a queen's contiguity matrix which considers neighborhood based on shared boundary. The weights are then standardized to ensure that weights attributed to each postal code sums up to one. Chapter 4 is a crash mobility and congestion analysis which focuses on congestion attributed to a crash event on a specific road section. It is a discrete choice outcome model for crash congestion severities. Finally, Chapter 5 discusses the overall conclusions and recommendations from the research findings.

1.4 Background of Geospatial Techniques in Road Safety Research

This section discusses examples of Geographic Information Science (GIS) techniques commonly used to analyze safety data. Mapping has been a primary role of GIS in the analysis. Maps show the crash location and are useful in visualizing clusters. However, as crash density increases, the amount of information that can be gathered from the map decreases due to point overlap. Therefore, mapping alone is insufficient to study the geospatial nature of crash events. The GIS tools used in this research are introduced in the following sections.

1.4.1 Global Moran's I Index

Global Moran's I is a measure of spatial autocorrelation based on feature location and attribute values. It is suitable for point features such as crashes with feature attributes, for example the number of crashes. It estimates whether features are clustered or dispersed. Local Moran's I identify statistically significant hotspots. It estimates the z-score and p-values which are a measure of statistical significance (Mitchell, 2005). A positive Moran's I Index shows that the dataset tends to cluster spatially while a negative index show that the points are dispersed. An Index value closer to zero indicate that the points are random.

1.4.2 Nearest Neighbor Analysis and Spatial Autocorrelation

Average Nearest Neighbor (ANN) analysis is used to test a null hypothesis that events/features are randomly distributed. The analysis returns a z-value and p-value which is a numerical approximation of the area under the curve for a known distribution (Mitchell, 2005). This technique is applicable if crash data is being analyzed as point data to test if crashes are random events or clustered events. Black (1991) and Levine, et al., (1995) used the nearest neighbor index to analyze crash data in Indiana and Honolulu respectively. The nearest neighbor index identifies if the observed feature distribution is clustered, random or dispersed by utilizing distance between features and the expected mean distance between the features. The same result can be achieved using cluster analysis and hotspot analysis or even kernel density estimation.

On the other hand, spatial autocorrelation measures the relationship between events based on both value and location. It evaluates whether the pattern expressed is either clustered, dispersed or random and calculates the Moran's I Index and the z-value and p-value to assess significance of the index (Mitchell, 2005). It is important in understanding spatial dependence and unobserved heterogeneity or patterns. A positive Moran's Index shows the dataset tend to

cluster spatially. For example, Erdogan (2009) used spatial autocorrelation analysis to show whether provinces with high rates of fatalities were clustered.

1.4.3 Kernel Density Estimation (KDE)

KDE estimates the density of features using a kernel function per unit area. Many researchers have used both planar and network kernel density estimation to analyses crash data. Flahaut, et al., (2003) compared the local spatial autocorrelation index and the kernel density estimation method to identify location and length of roads sections characterized by a concentration of accidents in Belgian (Flahaut, et al., 2003). Krishna et al (2005) used the kernel density estimation method to identify high pedestrian crash zones in Nevada. In addition, Anderson (2007) used KDE to investigate the merits of network analysis and area wide analysis in identification of road accident hotspots in London. Many other researchers have used KDE to analyze road safety data. For example, Pulugurtha, et al., (2007) used kernel method to identify high pedestrian crash zone. Borruso (2008) developed network density estimation and implemented in GIS then compared the method to planar kernel density estimation for cities in Italy and UK. Erdogan, et al., (2008) used kernel density method to identify hotspots in Turkey. Xie and Yan (2008) did a similar study by developing a network kernel estimation method and comparing with planar kernel density estimation. Anderson (2009) also used kernel density method to identify road accident hotspots in London.

1.4.4 Hotspot Analysis and Getis Ord Gi*

This tool calculates the G-i-star for each feature in the dataset and identifies statistically significant spatial clusters of high values (hot spots) and cold spots (Low values). It considers the feature values and the corresponding values of neighboring features (Mitchell, 2005). A feature

becomes a hotspot if it is statistically significant (p-value less than 0.05), has a high value and is surrounded by features with high values within a specified distance of analysis. A high z-score shows hotspot of high value features. A low z-score shows a cold spot of low value features (Mitchell, 2005). This method was successfully used to analyze spatial-temporal data on AIDS epidemic in san Francisco (Ord and Getis, 1995).

Hotspot analysis uses point data and creates a map of statistically significant hot and cold spots using the Getis Ord Gi* statistic values (Mitchell, 2005). Steil and Parrish (2009) developed a hotspot identification taxonomy and implemented in GIS by analyzing events on road segments and countermeasure activity on the segment. Finally, Gundogdu (2010) also used hotspot analysis to analyze crash data in Konya, Turkey. Finally, Khan et al., (2008) analyzed weather related crash patterns in Wisconsin using Getis Ord Gi*.

1.4.5 Geographically Weighted Regression (GWR) Analysis

Geographically Weighted Regression (GWR) is a linear formulation which models spatially varying features based on independent and dependent variables (Fotheringham, et al., 2000). It improves on standard regression by accounting for spatial autocorrelation. This study explores the application of Geographically Weighted Poisson Regression (GWPR) in analysis of crash data. GWPR is a combination of GWR and Poisson regression.

1.5 Background of Crash Congestion Analysis

The second aspect of this research focuses on the relationship between congestion severity and road accidents. Whereas it is expected that any random crash event will result into traffic congestion of some magnitude, it is important to quantify the severity of the resulting congestion. This can either be a low, medium or high severity congestion depending on the

impact on free flow speed and the total queue length of the affected vehicles. The relationship between congestion and accidents has been widely researched (e.g., Dickerson, et al., 2000; Quddus, et al., 2010; Hojati, et al., 2013; Hojati, et al., 2014). Whereas these research efforts show the impact of traffic congestion on accidents, the opposite is still a subject for further investigation. As such, there is a need to establish how much congestion is expected from a random crash event. The findings from these efforts is important in that they can help to minimize congestion while at the same time improve road safety. It is also inherent that a queue on an interstate is likely to result in a secondary crash at the tail of the queue. As such, these research efforts can also help mitigate the occurrence of secondary crashes. Though still, occurrences of secondary crashes can be a subject of independent investigation. Chapter 4 presents a detailed analysis and discussion of a crash mobility analysis which defines the relationship between congestion severity and road accidents.

CHAPTER 2. GEOGRAPHICALLY WEIGHTED POISSON REGRESSION ANALYSIS OF SAFETY DATA AND MACRO LEVEL SOCIOECONOMIC FACTORS – A CASE STUDY OF ALABAMA DUI CRASHES

2.1 Introduction

The need to integrate safety in infrastructure planning has been widely accepted as a step towards improving road safety. Some researchers introduced the aggregated crash prediction process as a "crash generation" concept (Naderan and Shahi, 2010). Some studies have indicated that this can be done at the macro or micro level (e.g., Lovegrove and Sayed, 2006; Huang, et al., 2016; Gomes, et al., 2017). An early attempt was made to estimate a macro level accident prediction model for planning purposes in Toronto using socioeconomic, traffic demand and network data (Hadayeghi, et al., 2002). Amoh-Gyimah (2017) gave an example of developing a crash count or crash rate model at macro level using socioeconomic, demographic, land use and transport network data (Amoh-Gyimah, et al., 2017). Previous studies have sought to establish a relationship between crash frequencies and socioeconomic factors (e.g., Hadayeghi, et al., 2010; Li, et al., 2013; Shariat-Mohaymany, et al., 2015; Amoh-Gyimah, et al., 2017). However, most techniques which have been used for planning purposes have included models estimated using generalized linear modeling methods (Hadayeghi, et al., 2010). The primary limitation of this method is that it assumes that the independent variable estimates are fixed "globally" throughout the entire planning region. This ignores inter- and intra-geographical variability among different regions.

This stationarity assumptions leads to violation of the Gauss-Markov requirements for absence of serial autocorrelation, multicollinearity and correlation among the error terms

(Wooldridge, 2013). Efforts have been made to address this problem and few examples are discussed in this section. Li (2013) used Geographically Weighted Poisson Regression (GWPR) to estimate a spatially varying model for county level crash data and compared it with a traditional Generalized Linear Model (GLM). Similarly, other researchers have investigated the relationship between the number of zonal collisions and potential transportation planning predictors, using the GWPR (e.g., Hadayeghi, et al., 2010). In addition, research has also been done on the relationship between crashes and socioeconomic factors (e.g., Abdel-Aty, et al., 2011).

Using this background, this research applies GWPR technique to explore and analyze driving under the influence (DUI) crashes in the State of Alabama. Alcohol-impaired driving accounts for substantial proportion of traffic-related fatalities in the United States (MacLeod, et al., 2015), and laws have been enacted in the United States to reduce alcohol related crashes (Romano, et al., 2015). This research focuses on road safety planning by analyzing the relationship between socioeconomic factors and accidents attributed to DUI. While it is acknowledged that safety can be enforced by the police through arrest, there are studies which have shown that DUI arrests alone cannot be sufficient as a countermeasure for drunk driving (Dula, et al., 2007). As such, it is important to also focus on education and awareness campaigns.

The broad objective of this study is to understand how macro level socioeconomic factors affect traffic crashes. This is achieved by undertaking a case study and developing a macro level crash prediction model for DUI crashes at postal code level and examining the relationship between DUI crashes and socioeconomic factors. To achieve this objective, a GWPR model for DUI crashes at the postal code level was estimated to define the relationship between DUI

crashes and socioeconomic factors while taking care of spatial dependence and geospatial autocorrelation among various postal codes.

2.2 Approach and Methodology

This section discusses the approach and methodology used in this study. To begin with, the Gauss-Markov assumptions are considered as listed below (Wooldridge, 2013):

- *Linearity* - linear relationship between the dependent variable and the parameters;
- *Zero error mean* - the expected mean of the error term is zero;
- *Homoscedasticity* - the error term has a constant variance;
- *No serial correlation* - the errors are not correlated over time and that there is no correlation among the dependent variable;
- *Deterministic parameters* - no correlation between the parameters and the error term; and
- *Multicollinearity* - the parameters are independent of each other and there is no multicollinearity.

Regarding statistical and econometric models based on crash data, it has always been assumed that there is no spatial dependence or correlation among the dependent variable. This oversight inherently violates the Gauss-Markov assumption for "*No serial correlation"*. In addition, due to presence of spatial heterogeneity, most statistical and econometric models violate the Gauss-Markov assumption for "*Deterministic parameters*" and "*Multicollinearity*" (Lesage, 1999; Lesage, 2008). Spatial dependence and spatial heterogeneity introduces problems because of spatial aggregation, spatial externalities and spill-over effects (Anselin, 1988) which

is at the core of regional science and geography as expressed by Tobler (1970) that *ceteris paribus:*

"everything is related to everything else, but near things are more related than distance things"

From a modeling perspective, spatial dependence simply means that an observation at location i depends on other observations at locations j where $i \neq j$ and can be expressed as follows:

$$y_i = f(y_i), i = 1 \dots \dots \dots n, i \neq j \quad (2.1)$$

Mannering and Bhat (2014) emphasized that issues related to unobserved heterogeneity, endogeneity, spatial and temporal correlations remain a huge methodological barrier in the statistical analysis of crash data and in the understanding of factors that affect the likelihood of road crashes (Mannering and Bhat, 2014).

2.2.1 Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a technique that allows parameter estimates to vary over different geographic regions (Li, et al., 2013). To do this, it incorporates coordinates x-y into the model. (Feuillet, et al., 2015). The general form of GWR equation is given as (Nakaya, et al., 2016):

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)\, x_{k,i} + \varepsilon_i, \quad (2.2)$$

Where:

$y_i, x_{k,i}\ and\ \varepsilon_i$ are dependent variable, independent variable and error term respectively at location i.

(u_i, v_i) represent the geographical location (co-ordinates) of location i

$\beta_k(u_i, v_i)$ represent the parameter estimate for location i for independent variable k.

GWR is better than traditional modeling techniques because it allows for varying parameter estimates and as such, may lead to new and different interpretation of the results. This research will particularly use the framework of a Geographically Weighted Poisson Regression (GWPR) technique which is generally preferred for modeling count data (Hadayeghi, et al., 2010; Aguero-Valverde, 2013; Shariat-Mohaymany, et al., 2015; Xu and Huang, 2015; Amoh-Gyimah, et al., 2017) as discussed in the following section.

GWPR was developed in 2005 (Nakaya, et al., 2005). It includes an advanced semi-parametric component which has been applied to investigate spatial heterogeneity in regional safety modeling and explore spatially structured varying relationships in Florida (Xu and Huang, 2015). The semiparametric version of GWPR is appropriate for regional crash modeling because it accounts for spatial correlation among crash data and independent variables by allowing for both local and global variables. Research has indicated that there is an important need to understand how variation in spatial units affects spatial heterogeneity (Amoh-Gyimah, et al., 2017). When compared to random parameter negative binomial regression, the semiparametric GWPR gives the same significant parameters with the same signs across spatial units but with varying magnitude of the coefficient which emphasizes the importance of accurately assessing the impact of spatial heterogeneity on the dependent variables. A study done in Costa Rica confirmed that spatial models are better at predicting crash frequencies (Aguero-Valverde, 2013). Again, GWPR was employed in Mashad, Iran to investigate the relationship between crash data and socioeconomic characteristics at Traffic Analysis Zone (TAZ) level (Shariat-Mohaymany, et al., 2015) which confirmed that GWPR performs better compared to the traditional GLM.

GWPR captures spatial non-stationarity. As mentioned above, it can take two formats, semiparametric and non-semiparametric. Poisson regression is used for count data modeling in which the dependent variable is either an integer greater than zero or zero. The two variants of GWPR are presented below (Nakaya, et al., 2016):

GWPR: $y_i \sim Poisson\left[N_i\ exp\left(\sum_k \beta_k(u_i, v_i)\ x_{k,i} + \varepsilon_i,\right)\right]$ (2.3)

Semiparametric GWPR: $y_i \sim Poisson\left[N_i\ exp\left(\sum_k \beta_k(u_i, v_i)\ x_{k,i} + \sum_l \gamma_l\ z_{l,i} + \varepsilon_i,\right)\right]$ (2.4)

Where:

Z_i is the l_{th} independent variable for the associated fixed coefficient value γ_i,

N_i is an offset variable at location i and represent the expected value of the dependent variable, normally, in Poisson regression, this value defaults to one.

The semiparametric version has both global fixed variables (whose parameter estimates do not vary by geographic location) and local variables (with varying parameter estimates depending on the geographic location). This technique enhances the prediction performance of the estimated model. Further, to satisfy all the Gauss-Markov assumptions, GWPR allows for standardization of all variables using the z-transformation (Nakaya, et al., 2016).

2.2.2 Kernel Functions

In this approach, a kernel function is used to estimate geographic weights based on Euclidian distance and bandwidth. The Euclidian distance is the distance between the geographic location of the dependent variable and geographic location of the independent variable. Therefore, the weight varies such that variables closer to the dependent variable are given higher weights than variables further away from the dependent variable. There are four different options for kernel type functions and weights estimation as shown below (Nakaya, et al., 2016):

- Fixed Gaussian: $$w_{ij} = exp\left(-d_{ij}^2 \Big/ \theta^2\right) \quad (2.5)$$

- Fixed bi-square: $$w_{ij} = \begin{cases} \left(1 - d_{ij}^2 \Big/ \theta^2\right)^2 & d_{ij} < \theta \\ 0 & d_{ij} > \theta \end{cases} \quad (2.6)$$

- Adaptive bi-square: $$w_{ij} = \begin{cases} \left(1 - d_{ij}^2 \Big/ \theta_{i(k)}\right)^2 & d_{ij} < \theta_{i(k)} \\ 0 & d_{ij} > \theta_{i(k)} \end{cases} \quad (2.7)$$

- Adaptive Gaussian: $$w_{ij} = exp\left(-d_{ij}^2 \Big/ \theta_{i(k)}^2\right) \quad (2.8)$$

Where:

$i\ and\ j$ are the regression and location indices respectively.

w_{ij} is the estimated geographical weight for independent variable at location j with reference to a dependent variable at location i.

d_{ij} is the Euclidian distance between $i\ and\ j$.

θ is a fixed bandwidth over the geographic area.

$\theta_{i(k)}$ represent an adaptive bandwidth defined as k_{th} nearest neighbor.

In this research, the adaptive bi-square kernel weight function was used because it can iteratively estimate the optimum bandwidth that fits the best model.

The GWPR model is calibrated by a kernel regression method which uses the distance based weighting scheme. Based on the adaptive bi-square kernel function, the bandwidth changes for different locations which as a result change the number of neighborhoods associated with a given dependent variable at location i. Thus, each estimated model for the ith dependent

variable has a unique optimum number of neighborhoods. In this approach, the best estimate of bandwidth is adopted as that which minimizes the Akaike Information Criterion (AIC) value (Feuillet, et al., 2015). Figure 2.1 represents the geographic weighting computation and bandwidth identification. Generally, regarding the adaptive spatial kernel, bandwidth tends to be wider in sparse geographic neighborhoods and narrower in a dense geographic neighborhood.

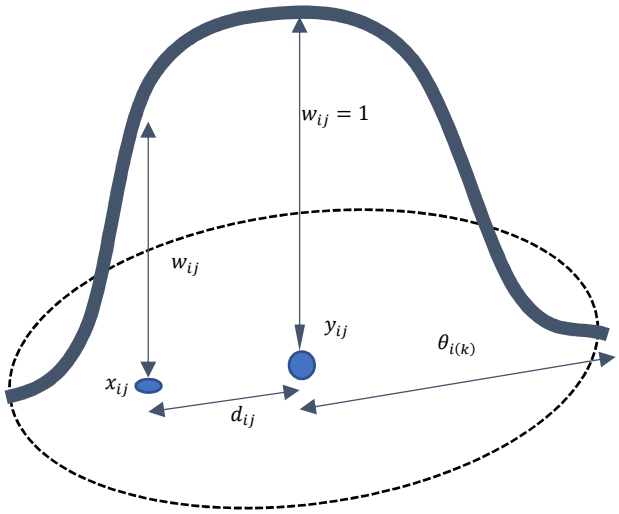


Figure 2.1: Representation of the geographic weighting estimation.

2.2.3 Goodness of Fit Statistics

To determine how good the data fits the model, four goodness of fit statistics will be estimated namely - (1) Mean Prediction Bias (MPB), (2) Mean Square Prediction Error (MSPE), (3) Mean Absolute Deviance (MAD), (4) Akaike Information Criterion (AIC). These measures have been used by other previous related research and are given as follows (Li, et al., 2013):

- Mean Prediction Bias (MPB)

$$MPB = \frac{(\sum_{i=1}^{n}(\hat{\mu}_1 - y_i)}{n} \tag{2.9}$$

MPB can be negative or positive. A positive value implies over estimation. While a negative value imply underestimation.

- Mean Square Prediction Error (MSPE)

$$MSPE = \frac{(\sum_{i=1}^{n}(\hat{\mu}_1 - y_i)^2)}{n} \quad (2.10)$$

A lower MSPE is preferred. Is also called mean absolute prediction error. Assesses the error associated with validation or external data.

- Mean Absolute Deviance (MAD)

$$MAD = \frac{(\sum_{i=1}^{n}|\hat{\mu}_1 - y_i|)}{n} \quad (2.11)$$

Smaller values of MAD are preferred. Gives average magnitude of variability of predictions.

- Akaike Information Criteria (AIC)

$$AIC = -2 * LL + 2 * (number\ of\ paramters) \quad (2.12)$$

AIC describes the trade off between bias and variance. The lower the AIC the better.

Where:

$\hat{\mu}_1$ represent the predicted estimates,

n represents the number of geographic regions,

LL denotes log-likelihood and

y_i represents the actual values of the dependent variable for each geographic region.

Finally, to check for serial autocorrelation, Moran's I statistics were estimated for the residuals of the prediction estimates. Moran's, I ranges from -1 to +1 and a 0 value indicates no spatial correlation among the residuals implying no serial autocorrelation (Mitchell, 2005).

2.2.4 Geographical Variability Test

In a GWPR, if the geographic covariation of the significant variables is not statistically significant, then the model defaults to a GLM. Therefore, to determine if the geographic covariation is statistically significant, a geographic variability test is done for each local significant variable. Noting that local variables are those whose coefficients vary from region to region (the parameter estimates are not fixed). The geographic variability test for variable k_i is done by comparing two models; one is the fitted GWPR model and the second is the model in which variable k_i is held constant while the other variables remain as fitted in the GWPR model. The comparison is done using the AIC output. A model with lower AIC is always considered to be better than a model with higher AIC (Gonzalez-Rivera, 2016). For example, to test for geographic variability of independent variable x_1 the two models are given as:

Fitted Model: $y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{1,i} + \varepsilon_i$ (2.13)

Fixed Slope Model: $y_i = \beta_0(u_i, v_i) + \beta_1 x_{1,i} + \varepsilon_i$ (2.14)

If the model with the fixed slope has a lower AIC than the fitted model, then the coefficient estimate $\beta(u_i, v_i)$ has a significant variation from region to region. The reverse is true if the fitted model has lower AIC than a fixed slope model. The two models are estimated using the same bandwidth (estimated using the adaptive bi-square kernel function). Finally, the difference in the models is shown by calculating the "Diff of Criterion" between the two models. If the fitted model has lower AIC then the "Diff of Criterion" will be positive implying there is no spatial variability. If the fitted model has a higher AIC, then the "Diff of Criterion" is negative implying there is significant geographic variability (Nakaya, et al., 2016).

2.3 Data

This study used two types of data sets - (1) crash data and (2) socioeconomic data. The details of each data type are described in the following section.

2.3.1 Crash Data

Crash data were obtained from Critical Analysis Reporting Environment (CARE) hosted by the Center for Advanced Public Safety (CAPS) at the University of Alabama. The data were for each crash record provided by the Alabama Department of Transport (ALDOT). Each crash record contains all details related to a crash recorded by the police at the time of the crash. The data analyzed included crash events which occurred from 2009 to 2013. For the current study, the data were filtered resulting in a total of 21,818 crash record where DUI was the primary contributing circumstance. Each crash record contained the postal code of "Driver 1" indicating the driver who was determined to be "at-fault" by the reporting police officer. The DUI crashes were then sorted by postal code. The mean DUI crash per postal code was determined to be 13.9 with a standard deviation of 25.11. Socioeconomic data were obtained from the US Census Bureau (U.S. Census Bureau, 2017). Population-based crash rates were computed by dividing the DUI crash frequencies by the population of residents in each postal code. A total of 781 postal codes in Alabama were analyzed and model estimation was based on 639 as these were the ones with complete data sets for all variables. Table 2.1 summarizes the relevant crash data and 46 socioeconomic data categories assembled for the study.

Table 2.1 Variables considered in Spatial Econometric Models

Variable Description (by postal code)	Mean (Std. Dev)	Variable Description (by postal code)	Mean (Std. Dev)
Crash rates normalized by population	0.01(0.01)	Ln (males between ages 15 to 17)	1.24(0.73)
% of population between ages 15 to 17	4.08(3.20)	Ln (males who are separated)	0.55(0.85)
% of females between ages 15 to 17	3.82(4.05)	Ln (population who are separated)	0.81(0.74)
% of males between ages 15 to 44	37.76(13.02)	Ln (Crash rates normalized by population)	-2.76(0.42)
% of females between ages 15 to 44	36.49(12.71)	Ln (population who are living in rented housing)	2.90(0.94)
% of population aged above 65	16.23(10.02)	Ln (% of females who have less than high school certificate)	2.09(1.43)
% of males aged above 65	14.28(9.89)	Ln (female population)	7.38(1.67)
% of females aged above 65	18.11(11.13)	Ln (% of male with bachelor's degree or higher)	0.64(1.03)
Employment rate	47.68(13.33)	Ln (% of male who have less than high school certificate)	2.46(1.43)
% of residents living in rented housing	24.43(16.58)	Ln (population of residents with less than high school education)	2.51(1.24)
% of population living in their own housing	74.19(18.54)	Ln (population who have bachelor's degree or higher)	0.97(1.10)
% of population who are married	49.56(15.69)	Ln (unemployment rate)	2.26(0.84)
% of population who are divorced	11.70(5.37)	Median income ($10,000)	4.02(1.79)
% of population who are never married	12.69(24.75)	Ln (divorced population who are black)	1.85(1.21)
% of males who are married	51.96(16.57)	Ln (divorced population who are white)	2.24(0.82)
% of males who are divorced	11.39(6.48)	Employment rate of population who are between ages 16 to 19	2.43(1.35)
% of females who are married	48.36(16.68)	Ln (male population)	7.33(1.68)
Median income	40227(17857.29)	Ln (population between ages 18 to 24)	2.00(0.77)
Employment rate for females between ages 20 to 64	54.85(16.14)	Ln (males between ages 18 to 24)	2.01(0.85)
Employment rate for males between ages 20 to 64	64.48(19.30)	Ln (females between ages 18 to 24)	1.88(0.83)
Average household size (persons/household)	2.55(0.39)	Ln (population between ages 15 to 44)	3.53(0.61)
% of female residents with bachelor's degree or higher	6.22(11.80)	Ln (female worker force)	6.25(1.88)
% of female with some college education	41.96(26.09)	Ln (male work force)	6.40(1.85)
% of all population with only high school certificate	31.85(18.95)	Ln (male who are never married)	0.71(0.97)
Employment rate of population who are between ages 20 to 24	50.86(26.55)	Ln (females who are divorced)	2.26(0.79)
Total population	7498.73(8801.94)	Ln (females who are never married)	0.80(0.99)
Ln (% of females with only high school certificate)	2.70(1.43)	Ln (males with some college education)	2.89(1.42)
Ln (% of females who are separated)	0.88(0.84)	Ln (residents with some college education)	3.25(1.19)
Ln (males with only high school certificate)	3.03(1.35)		

Ln: Natural logarithm, %: Percentage

A histogram of crash frequencies in each postal code in Alabama is shown in Figure 2.2 below. From Figure 2.2, the histogram for crash frequencies in Alabama postal codes is right skewed. Most of the postal codes had DUI crash frequencies below 50 over the five-year period which implies that on average each postal code had about 10 drivers involved in DUI crashes annually.

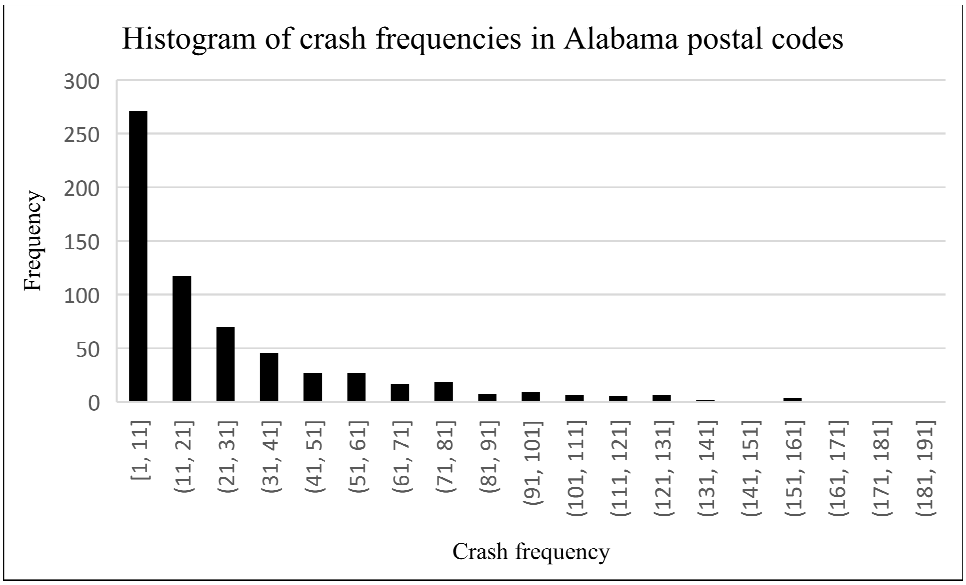


Figure 2.2: Histogram of crash frequencies in Alabama postal codes

A general distribution of the DUI crash frequencies in Alabama is shown in the line plot below (Figure 2.3). The line plot shows that the crash frequencies are stationery and as such can be modeled without transformation (Gonzalez-Rivera, 2016). This research focused on the postal code of the drivers that caused the accidents. GWPR regression technique is applied because the dependent variable is count data taken as crash frequency. A spatial distribution and cluster analysis of the crash frequencies in each postal code is presented in Figure 2.4.

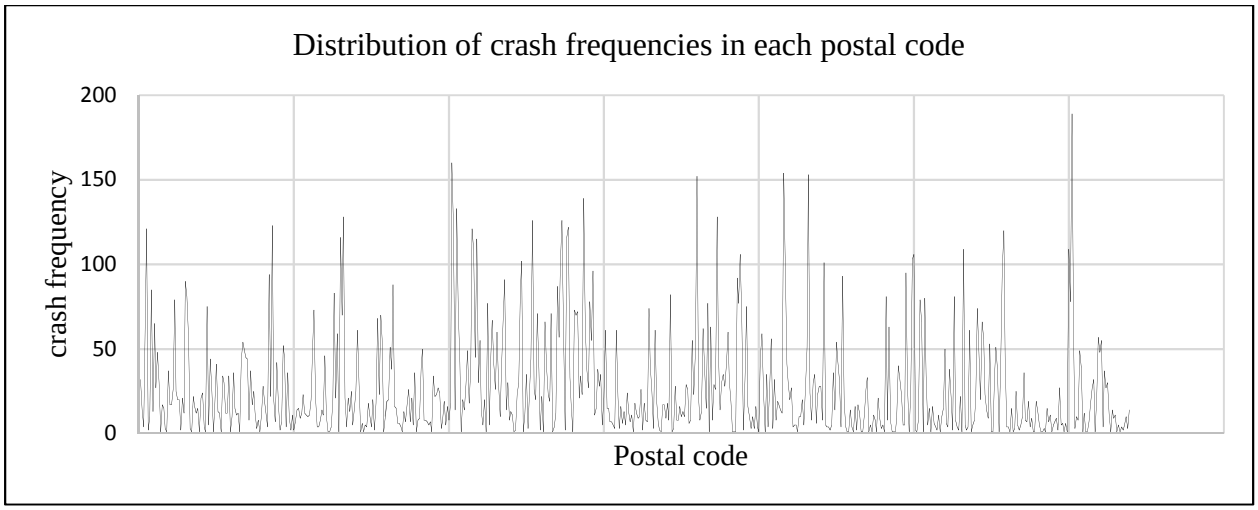
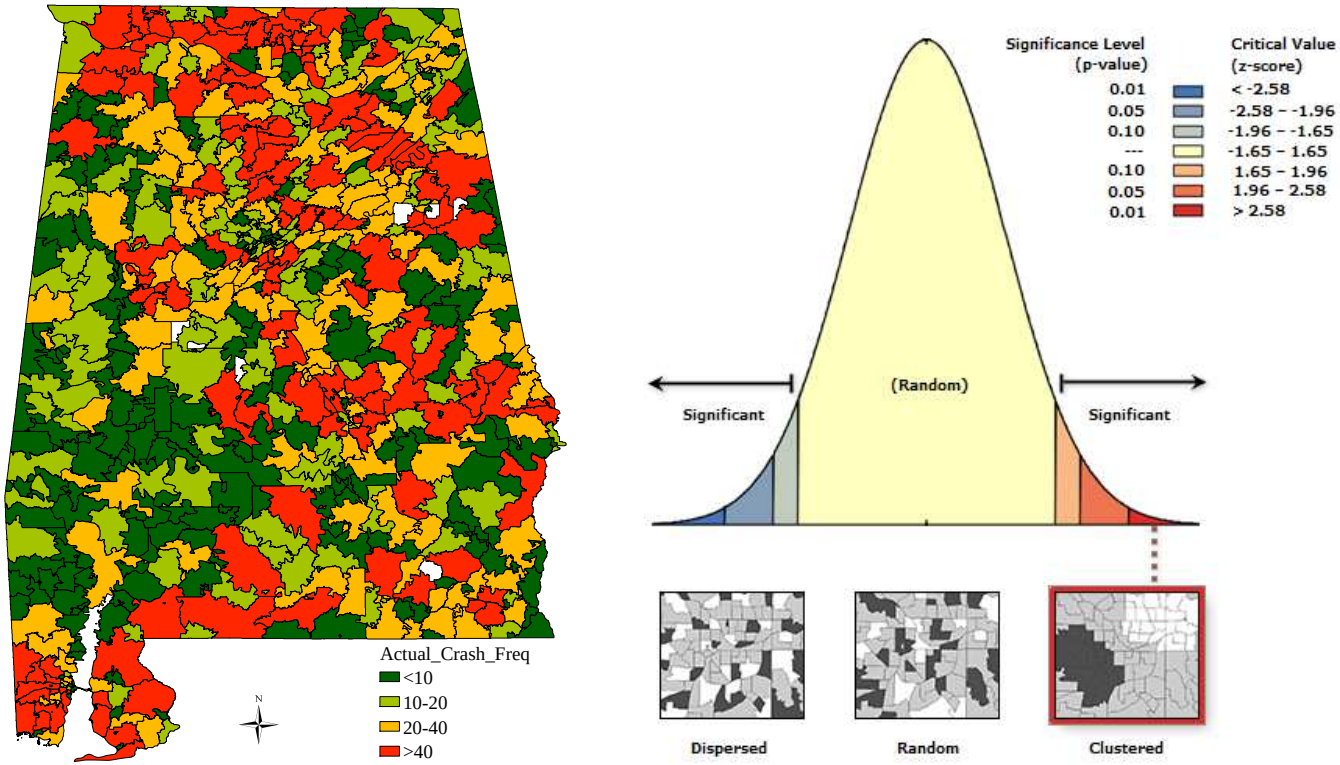


Figure 2.3: A general distribution of crash frequencies in each postal code



(a) (b)

Figure 2.4: Distribution of drivers causing DUI crashes in each postal code in Alabama.

From Figure 2.4 (b) the estimates of the global Moran's I indicate a z-score of 8.41. As such, there is a less than 1% likelihood that this clustered pattern could be the result of random chance. The estimated Moran's I index is 0.185 and a p-value of 0.000 for the actual crash frequencies. Table 2.2 shows the summary of the Moran's I index.

Table 2.2: Global Moran's I Summary

Moran's Index:	0.185
Expected Index:	-0.002
Variance:	0.001
z-score:	8.410
p-value:	0.000

Further, to understand the DUI crash patterns, a cluster analysis was done as shown in Figure 2.5 using a Getis-Ord Gi* hotspot analysis of the crash data. It depicts the high-high clusters areas which shows the postal codes where high crash frequencies are neighbors to each other and are statistically significantly clustered. High-low clusters indicate areas where postal codes with high crash frequencies are neighbors with postal codes with low crash frequency and they are significantly clustered. Low-high clusters indicate areas where postal codes with Low crash frequencies are neighbors with postal codes with high crash frequency and they are significantly clustered. And low-low clusters indicate areas where postal codes with low crash frequencies are neighbors with postal codes with other low crash frequency and they are significantly clustered.

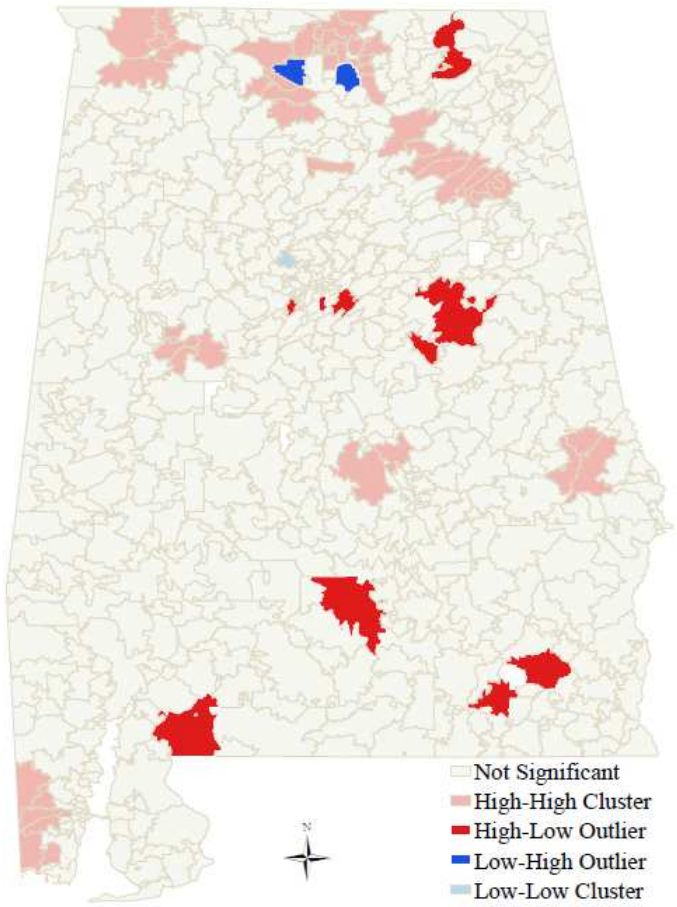


Figure 2.5: Cluster analysis of DUI crash frequencies.

2.3.2 Socioeconomic Data

Socioeconomic data were obtained from the US census website (United States Census Bureau, 2017). The data were filtered and cleaned for each postal code in Alabama. A total of 46 postal code related parameters were considered and matched with DUI crashes. Table 2.3 presents a summary of the significant independent variables. For comparison purposes, all socioeconomic variables considered in the analysis are summarized in Table 2.1 (see section 2.3.1).

Table 2.3: Variable descriptive statistics per postal code

Variable	Mean	Standard deviation
Employment rate	47.604	13.464
Percent of people living in rented housing	24.587	16.820
Income ($10,000)	4.022	1.788
Population density	302.121	729.960

2.4 Model Estimation Results

A non-semiparametric GWPR model was estimated using GWR4.09 (Nakaya, et al., 2016). This section discusses the estimation results. The variables identified have a significant geographical variability across the State of Alabama. As seen in Table 2.4, the significant variables include employment rate, percentage of people living in rented housing, income and population density.

The estimates are given as a range of values depicting geographical variability and demonstrating that the coefficients vary spatially across the State. The significant variables were taken as those that produced significant parameters in at least 80% of the postal codes. This percentage is acceptable where estimates vary geographically and has been used in other studies, (e.g., Li, et al., 2013). The estimates include the minimum, lower quartile, mean, median, maximum, and upper quartile value of the coefficient as presented in Table 2.4.

Table 2.4: Results for GWPR by postal code (Local independent variable estimates)

Variable	Minimum	Lower quartile	Mean	median	Maximum	Upper quartile
Intercept	0.000	3.009	3.313	3.356	10.491	3.780
Employment rate	-0.935	0.000	0.316	0.333	1.499	0.629
Percent of people living in rented housing	-0.747	0.000	0.343	0.268	1.703	0.672
Income ($10,000)	-1.122	-0.235	0.039	0.000	1.696	0.269
Population density	-1.845	-0.081	1.093	0.041	21.481	1.207
MAD			13.537			
MSPE			424.871			
% deviance explained			0.622			
Moran's, I of residuals			-0.030			
n			639			

The constant term has a minimum value of zero and a maximum value of 10.5 with the mean and median value tying at 3.3. This implies that, all factors being constant, a postal code can have no resident driver being involved in DUI crash or up to a maximum of 11 resident drivers getting involved in a DUI crash over a period of five years. This makes sense given that the dependent variable is a non-negative count data.

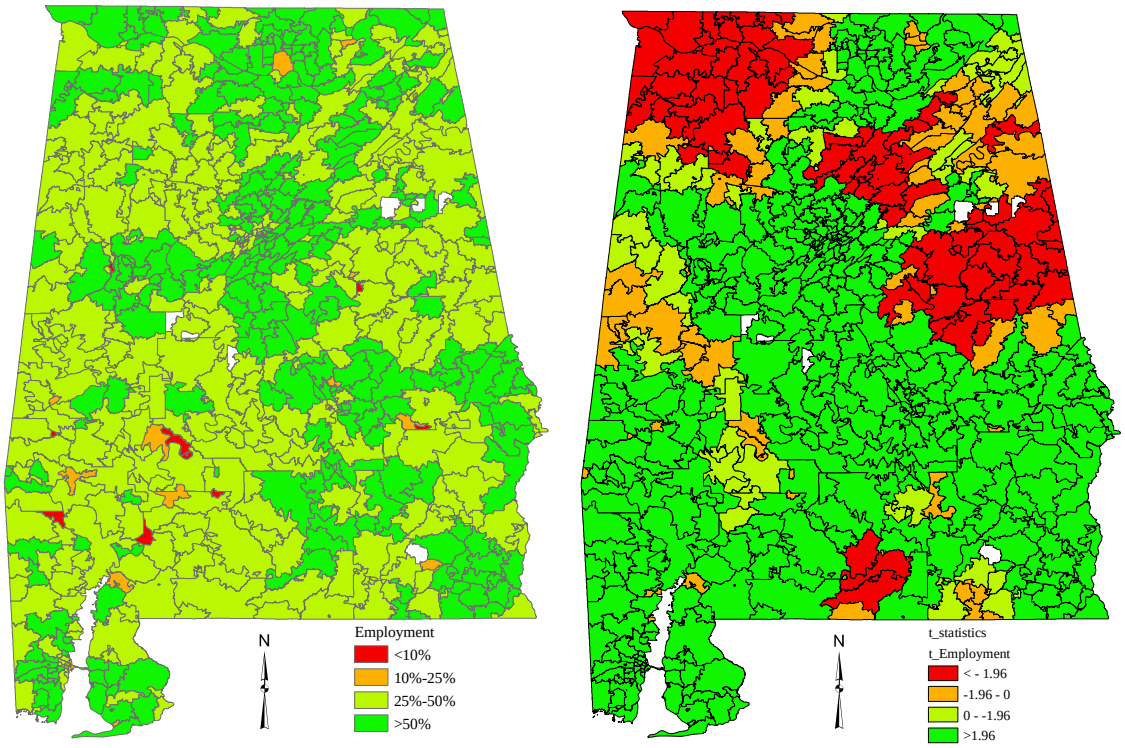
A detailed procedure of how the significant variables were identified is included in Appendix 2A. The following section examines and discusses the spatial distribution of each significant variable.

2.4.1 Employment

Employment rate in this analysis represent the percentage of labor force that is in gainful employment and is estimated as the number of employed persons as a percentage of total labor force per postal code. From Table 2.4, the estimated coefficient for employment rate has a minimum value of -0.9, a mean of 0.3 and a maximum value of 1.5. Therefore, it can be generalized that employment rate has a positive relationship with DUI crashes except for some rural areas where the relationship is negative.

Figure 2.6 (a) presents the spatial distribution of employment rate in Alabama in each postal code. There are few postal codes with less than 10% in formal employment. From the map, about 45% of the postal codes have employment rate of between 25%-50%. The remaining 45% of the postal codes have employment rate greater than 50%. Figure 2.6 (b) presents the corresponding t-statistics. The four clusters in red shows the postal codes where an increase in employment is negatively associated with DUI crashes. While the green areas show the postal codes where an increase in employment is positively associated with DUI crashes. The regions where DUI crash exhibit a negative association with employment are predominantly in rural areas of Alabama. As such, Figure 2.6 (b) indicates that an increase in employment in the rural areas reduce DUI crashes. And as observed, even though some rural areas have high employment rate, a further increase in employment potentially reduce DUI crashes. These findings can be investigated further for example, - what is the effect of dry and wet counties on DUI crash frequency given the dynamics of employment rate? An opposite relationship is observed in all urban areas where an increase in employment increases DUI crashes. This can also be a function of unobserved heterogeneity in the cosmopolitan urban areas. It is also a subject for further investigation considering factors such as changes in urban mobility given the dynamics in

employment rate. It is also seen (from Figure 2.6 (b)) that there are postal codes (about 10%) where employment is insignificant – these areas are all in the rural areas.



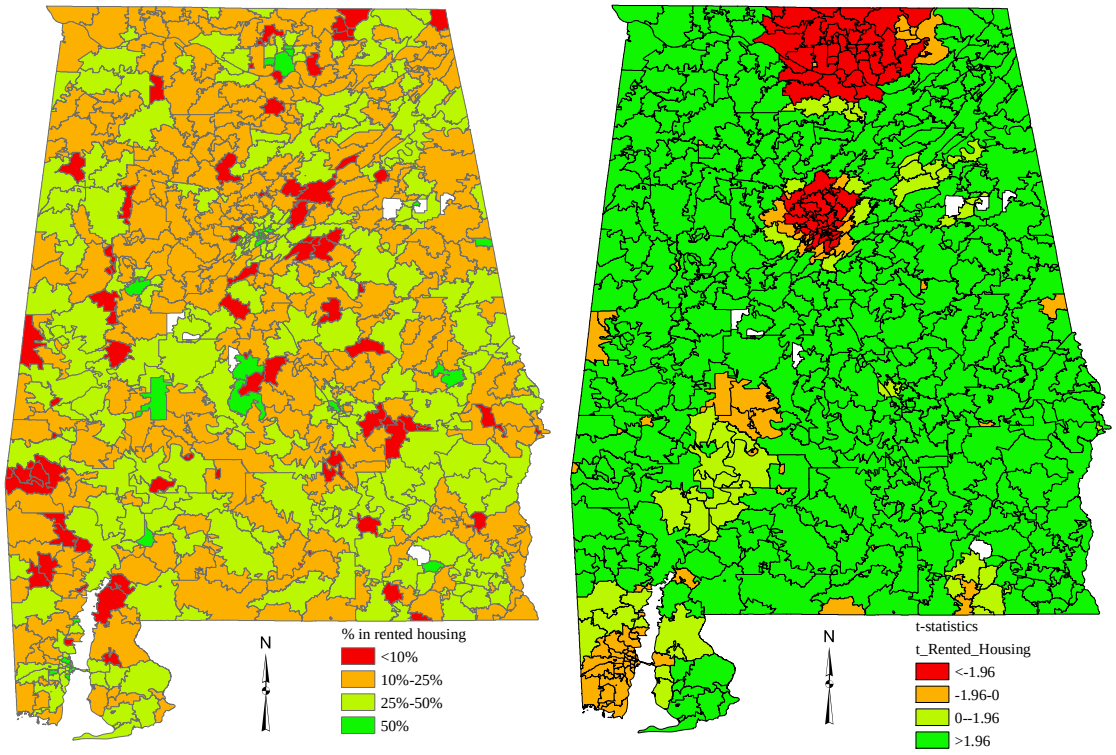
(a) (b)

Figure 2.6: Distribution of employment rate and corresponding t-statistics

These findings concur with a study done by Huang et al., (2010) in Florida which showed that, generally, counties with higher levels of unemployment exhibit lower overall crash risks. It is also consistent with the research done by Karl et al (2006) in Hawaii which showed that vehicle-to-vehicle crashes were associated with high employment.

2.4.2 Housing Characteristics

From Table 2.4, the coefficient for the percent of people living in rented housing has a minimum value of -0.7 and a maximum value of 1.7. The mean and median are both 1.3. Figure 2.7 (a) shows the spatial distribution of the percentage of people living in rented housing. Figure 2.7 (b) shows the t-statistics for percent of people living in rented housing in each postal code.



(a) (b)

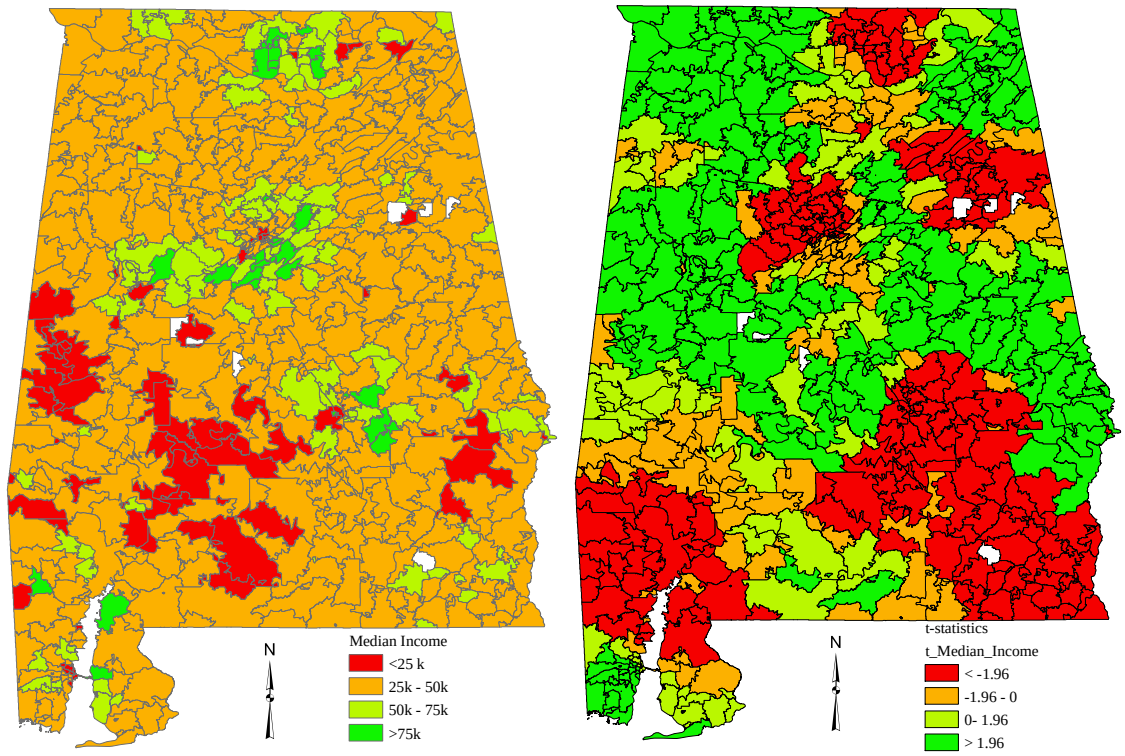
Figure 2.7: Distribution of people living in rented housing and corresponding t-statistics.

Part (a) of Figure 2.7 indicate that most postal codes have between 10% to 50% of the residents living in rented housing. Generally, the postal codes where most people live in rented housing are in urban areas while postal codes where few people live in rented housing are in rural areas. From Figure 2.7 (b), areas in green indicate postal codes where an increase in percent of residents living in rented housing lead to an increase in DUI crashes. While areas in red indicate postal codes where, as the percent of residents living in rented housing rise, DUI crashes decline. Owning a home can be associated with improved living standards (Nakaya T. et al., 2005). As such, it is plausible that when living standards improve in rural regions, then DUI crashes decline. However, in the metropolitan areas, for example, Jefferson county, Madison county and parts of Limestone county, in this study, results indicate that as rate of urbanization rise, DUI crashes decline. Generally, these urban areas are mostly dominated by the population

who are highly educated and who have a high perception of permanent income. The maps show that, in general, rental housing is positively associated with DUI crashes except for the urban/educated regions (Birmingham and Huntsville areas) whose perception of permanent income is relatively high. This explains why an increase in rental housing is negatively associated with DUI in the urban/educated regions (Chirinko and Harper, 1993). Overall, the result indicates that there is a positive relationship between rental housing and DUI crashes in rural areas but not in urban areas.

2.4.3 Income

Figure 2.8 (a) present the spatial distribution median income in Alabama postal codes. As seen, most postal codes have average annual median income between 25,000 to 50,000. Figure 2.8 (b) present the corresponding t-statistics for the median income. The areas in green indicate the postal codes where income has a positive relationship with DUI crashes. Most of the green areas are in rural Alabama. Interestingly, the positive and negative association between income and DUI crash frequencies share a 50-50 split. This is not paradoxical noting that the relationship between income and crashes has been widely investigated (e.g., Chirinko and Harper, 1993; Li, et al., 2013; Rhee, et al., 2016; Amoh-Gyimah, et al., 2017).



(a) (b)

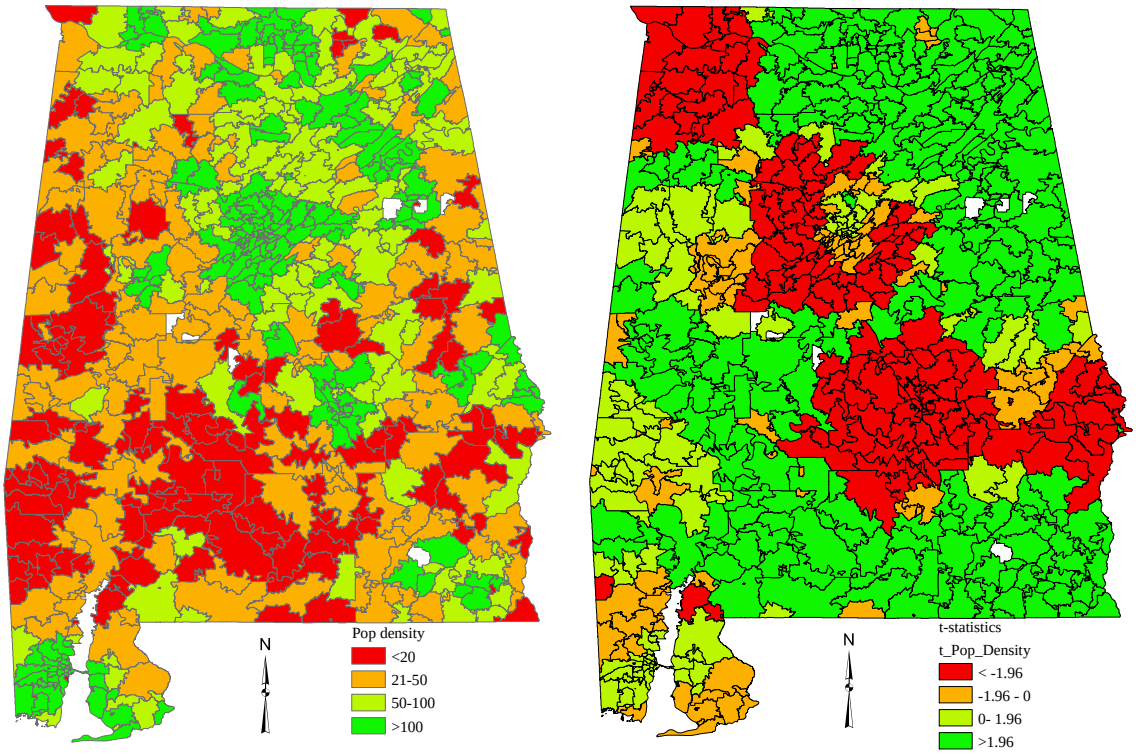
Figure 2.8: Distribution of median income and corresponding t-statistics

On the other hand, the areas in red in Figure 2.8 (b) represent the postal codes where there is a negative association between DUI crashes and income. The results show that there is a significant variation on the effect of income on DUI crashes. This is probably a function of unobserved heterogeneities and a subject for further investigation. The effect of income on DUI crashes in Alabama is significantly non-stationery. The positive association can be attributed to an increase in consumption expenditure as depicted by Chirinko and Harper (1993). This also agrees with other studies (e.g., Li, et al., 2013; Rhee, et al., 2016; Amoh-Gyimah, et al., 2017). On the other hand, the negative association points in the direction of other research findings that high-income level has a protective influence on alcohol related crashes (for example Romano, et al., 2006). A study by Romley, et al., (2007) also confirmed that density of liquor stores is higher in low income neighborhoods than in high income neighborhood which, increases the exposure

level of residents living in low income postal codes. As a result, when income rise in the rural neighborhood, the residents are more likely to indulge in increased leisure consumption such as alcohol.

2.4.4 Population Density

Figure 2.9 (a) presents the spatial distribution of population density in Alabama per postal codes. Figure 2.9 (b) shows the corresponding t-statistics for population density. Results of the GWPR showed that population density significantly influence DUI crash frequencies per postal code. The effect significantly varies spatially as demonstrated in part (b) of the map. The areas in green represent the postal code where an increase in population density leads to an increase in DUI crashes. The areas in red are those where an increase in population density leads to a decline in DUI crashes. These are mostly urban and cosmopolitan areas. Despite population density being high in urban areas (a), further increase in the density makes DUI crashes to decline. While density is low in rural areas (a), however, if the density increase, then DUI crashes also rise in rural Alabama.



(a) (b)

Figure 2.9: Distribution of population density and corresponding t-statistics

Other studies have shown that when population density increase, crash rates are likely to increase (Hadayeghi, et al., 2010; Li, et al., 2013; Xu and Huang, 2015). This finding is quite intuitive noting that DUI crashes as an outcome is a function of population. From the maps, positive association is observed in rural areas while negative association is observed mostly in urban areas. This points to an underlying problem in rural areas in which the increasing population are probably picking up the DUI habit as a regional norm with the reverse being observed in urban areas.

2.4.5 Goodness-of-Fit (GoF) Statistics

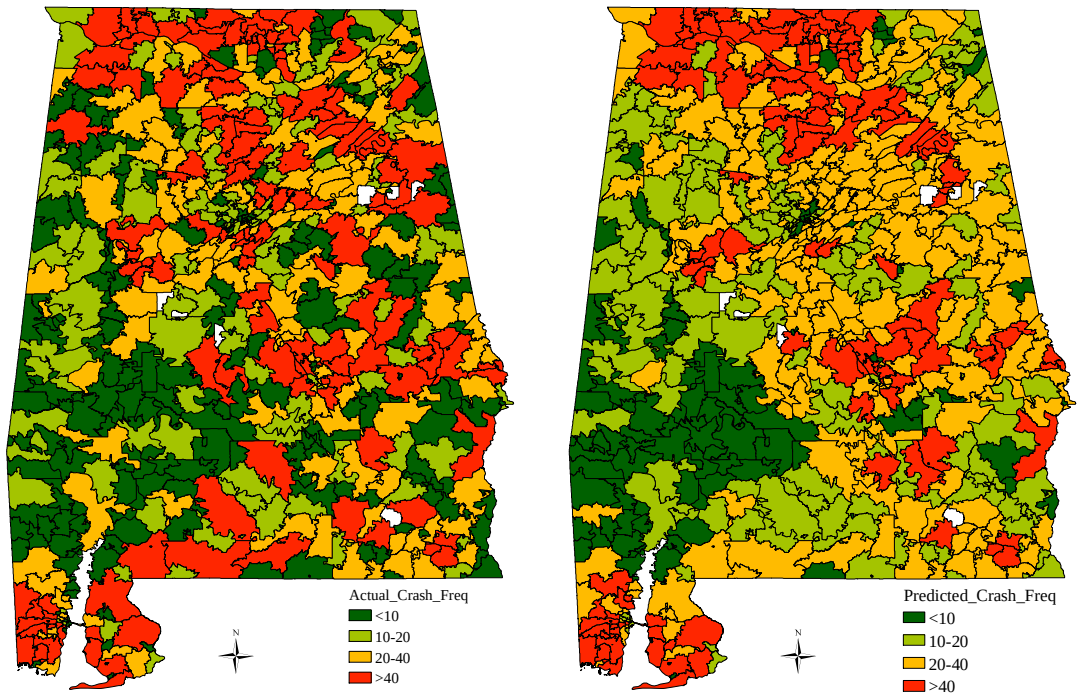
The GoF statistics are presented in Table 2.5. It shows that the percent deviance explained in about 62.2% while the mean absolute deviation is 13.54. While the mean square prediction error is 424.8. Most important is the Moran's I of the residual which indicates that

there is no spatial autocorrelation among the residuals which eliminates any serial autocorrelation.

Table 2.5: Goodness of fit statistics

MAD	13.537
MSPE	424.871
Percent deviance explained	0.622
Moran's, I of residual	-0.030
n	639

The fitness statistics in this study are used to appraise and comprehend the performance of the model. Figures 2.10 (a) and (b) shows the spatial distribution of the actual crashes versus the spatial distribution of the predicted crashes. The figures show that the estimated model is good at predicting the DUI crashes since the results are a fair representation of the true values.



(a) (b)

Figure 2.10: Distribution of actual versus predicted crash frequency per postal code

2.4.6 Testing for Serial Correlation

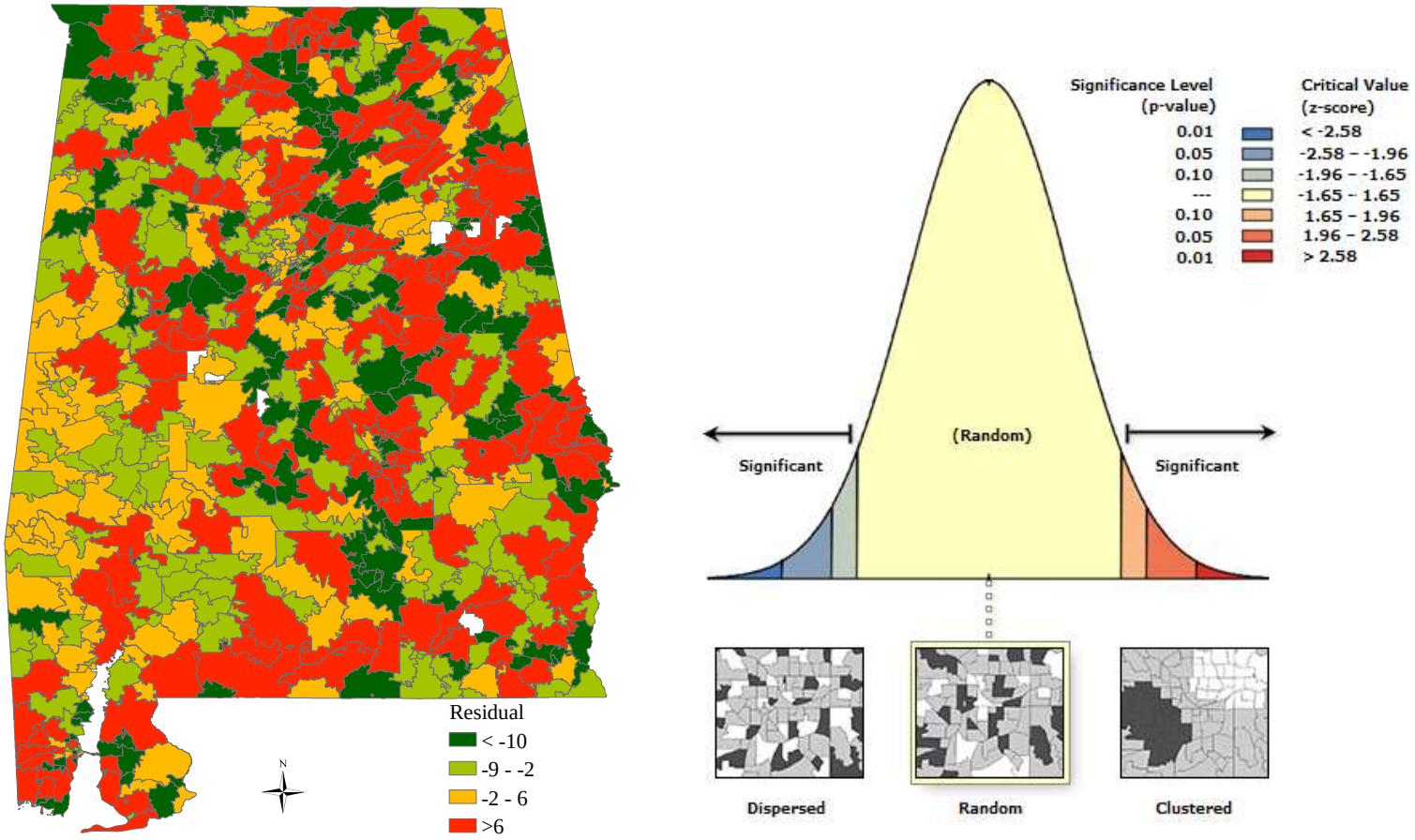
It is important to test for heteroskedasticity and know whether the error terms are random with a mean of zero. This can be done using the Breusch-Pegan test which is done by regressing the squared residuals against all significant variables and testing for significance of the variables. Alternatively, a White test or Breusch Godfrey test can also assess if there is serial autocorrelation (Wooldridge, 2013). Test for serial autocorrelation can also be achieved by estimating the global Moran' I index for spatial autocorrelation where parameters are geographically referenced. This is done by estimating the Local Indicator for Spatial Autocorrelation (LISA). Figure 2.11 (a) shows the spatial distribution of the residual terms while Figure 2.11 (b) shows the LISA for the residuals.

From the map of residuals and LISA analysis shown in Figure 2.11, the residuals indicate that the z-score value is -1.29, the pattern does not appear to be significantly different than

random. Summary of LISA analysis is shown below. The p-value is 0.197 showing that there is not enough evidence to indicate that the residuals are clustered which indicates absence of serial autocorrelation as required by the Gauss-Markov assumptions (Wooldridge, 2013).

Table 2.6: Global Moran's I Summary

Moran's Index:	-0.030
Expected Index:	-0.002
Variance:	0.001
z-score:	-1.290
p-value:	0.197



(a) (b)

Figure 2.11: Spatial distribution of residuals and Moran's I test for autocorrelation index

2.5 Geographic Variability Test Results

The results for geographical variability test for local coefficients is shown in Table 2.7. A test for geographical variability is performed on every significant variable to establish if the parameter is local or global. Global parameters are fixed and do not have geographical variability while local parameters significantly vary from region to region. From Table 2.7, the result of difference of criterion is negative for all variables which indicate that they significantly vary among the postal codes.

Table 2.7: Results of geographical variability test of local coefficients using chi-square test

Variable	Difference of deviance	Difference of degree of freedom	Difference of Criterion
Intercept	2801.351	23.1726	-2733.353
Employment rate	674.2021	23.451	-605.422
Percent of people living in rented housing	1255.2911	22.797	-1188.350
Median income in ten thousand	741.7401	19.863	-683.095
Population density	1140.628	17.596	-1088.455

2.6 Kernel Density Estimation (KDE)

KDE works by using a specified search radius and feature value that calculates the magnitude per unit kernel of a function (ESRI, 2017). It estimates the probability of a kernel function having a random variable outcome of the dependent variable. It is a technique that is used to study patterns and identify hotspots using GIS technology (Anderson, 2009). The density estimate for each location is given as follows (Fotheringham, et al., 2000):

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{d_i}{h}\right) \quad (2.15)$$

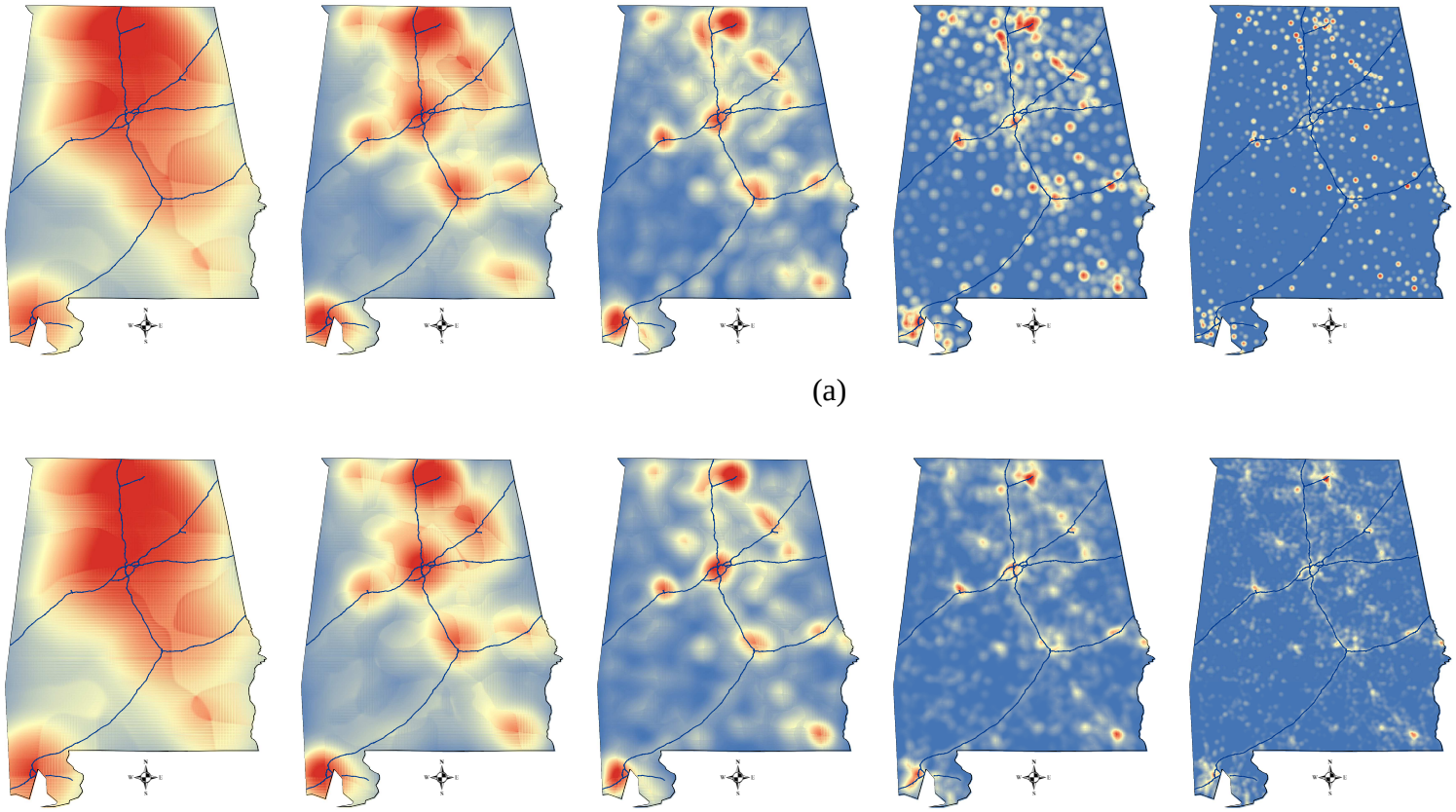
Where $f(x, y)$ is, the density estimates at location (x, y), n is the total number of observations, h is the bandwidth, K is the kernel function, $d_i$is the distance between location (x, y) and location of the ith observation. The K function takes care of the "distance decay effect" such that the longer the distance between a point and the reference location the less the point is weighted in calculating the overall density (Xie and Yan, 2008). Whereas this study will utilize the planar KDE technique, there are other methods that are more suitable for network based approach to estimating kernel density such as one developed by Okabe (Okabe, et al., 2006).

2.6.1 Overview

KDE technique is used to identify accident hotspots based on the assumption that the accident occurs in locations that are spatially dependent as such exhibit spatial interaction among the regions (Flahaut, et al., 2003). Anderson (2009) used kernel density estimation (KDE) to Identifying road accident hotspots in London. A kernel density estimation map was created and subsequently disaggregated by cell density to create a basic spatial unit of an accident hotspot. KDE has been used in several analyses of crash data (Xie and Yan, 2008; Anderson, 2009; Xie and Yan, 2013; Bíl, et al., 2013). The KDE method has two advantages: first, it spreads the risks by using a kernel around a certain band width, as such, it does not simply work with specific location of the accident. Second, it uses arbitrary spatial units which are homogenous across entire study area, it is very suitable where the spatial units of study have different sizes (Anderson, 2009).

2.6.2 Kernel Density Estimation Results

In this study KDE has been used to identify hotspots for two categories of data. First, are hotspots for drivers who get involved in DUI crashes. Second is the hotspots for locations where DUI crashes occur. This has been done at reducing bandwidth (search radius) from a maximum of 1.0 map units to 0.05 map units. The two corresponding maps for driver postal codes and crash locations is presented in Figure 2.12 (a) and (b). The hotspots are areas that can be identified for further investigation and analysis.



(a)

(b)

Figure 2.12: kernel density estimation for driver postal code and location of DUI crashes.

(a) KDE for postal code of drivers involved in DUI crash: Reducing search radius (1.0, 0.5, 0.25, 0.1, 0.05).
(b) KDE for location of DUI crashes: Reducing search radius (1.0, 0.5, 0.25, 0.1, 0.05).

2.7 Conclusion

In this research, DUI crashes were explored and analyzed by estimating a GWPR model to help understand the relationship between the DUI crashes and socioeconomic factors per region while taking care of spatial dependence among various postal codes. A KDE map was also produced to give more insight on the DUI hotspots. It is important to understand the reasons behind geographical variations in human behavior. Particularly, to explore how variation in socioeconomic factors influence behavior that contribute to road accidents. In this way, relevant policies and decisions can be made to mitigate to improve safety. Socioeconomic factors are vastly heterogenous across geography. And thus, is the relationship between aggregated socioeconomic factors at macro level and road accidents. For example, a rural neighborhood can have a series of other explanatory variables that effecting different driving behavior. The GWPR is a statistical tool which takes care of the extensive spatial heterogeneity and spatial non-stationarity. This technique is used in this study to demonstrate that the relationship between DUI crashes and socioeconomic factors is not the same everywhere in the State of Alabama. For example, the results suggest that the relationship between DUI crashes and the important factors namely -employment, income, population density and type of housing is not stationery but vary spatially within the State. The spatial variations are mapped in Figures 2.6 – 2.9.

This study increases knowledge and creates awareness on the local nature of the impact of socioeconomic inequalities on crashes, particularly, those attributed to DUI. This exposition will add value and further advance the understanding of the relationship between human behavior and road accidents. In respect of the unobserved heterogeneities, these local relationships can be examined further. It is also important to mention that some of the observation might be counterintuitive and this is not abnormal with GWPR models as observed

by a couple other previous researcher (for example Hadayeghi, et al., 2010; Amoh-Gyimah, et al., 2017).

Finally, whereas this study has been done at postal code level, it is important to consider existing planning level hierarchical boundaries as has been done by studies that applied traffic TAZ for safety planning (Abdel-Aty, et al., 2011). Other units of study already considered include census tracks, census wards, and census blocks (Amoh-Gyimah, et al., 2017). Some researchers have already estimated crash prediction models at TAZ level using land use data (Pulugurtha, et al., 2013).

2.8 References

Abdel-Aty, M., Lee, J., Siddiqui, C. and Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A, Volume 49, pp. 62-75.

Abdel-Aty, M., Siddiqui, C., Huang. H. and Wang, X., 2011. Integrating Trip and Roadway Characteristics to Manage Safety in Traffic Analysis Zones. Transportation Research Record: Journal of the Transportation Research Board, Issue 2213.

Aguero-Valverde, J., 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accident Analysis and Prevention, Volume 59, pp. 365-373.

Alabama Department of Transport (ALDOT), 2014. Drive Safe Alabama. [Online] Available at: http://drivesafealabama.org/uploads/files/ALDOT-2014CrashFactsBook.pdf [Accessed 21 09 2017].

Amoh-Gyimah, R., Saberi, M. and Sarvi, M., 2017. The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Analytic Methods in Accident Research, Volume 13, pp. 28-51.

Anderson, T., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis and Prevention, Volume 41, pp. 359-364.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Department of Geography and Economics, University of California, Santa Barbara: Kluwer Academic Publishers.

Bill, M., Andráši, R. and Janoška, k., 2013. Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accident Analysis and Prevention, Volume 55, pp. 265-273.

Chirinko R. and Harper E., 1993. Buckle up or slow down? New estimates of offsetting behavior and their implications for automobile safety regulation. Journal of Policy Analysis and Management, 12(1), pp. 270-296.

Chaloupka, F., Saffer, H. and Grossman, M., 1993. Alcohol-Control Policies and Motor-Vehicle Fatalities. The Journal of Legal Studies, 22(1).

ESRI Press, 2017. Incremental Spatial Autocorrelation. [Online] Available at: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/incremental-spatial-autocorrelation.htm [Accessed 09 Aug 2017].

Feuillet, T., Charreire H., Menai M., Salze P., Simon C., Dugas J., Hercberg S., Andreeva V., Enaux C., Weber C., and Oppert J. 2015. Spatial heterogeneity of the relationships between environmental characteristics and active commuting: towards a locally varying social ecological model. International Journal of Health Geography, 14(12).

Flahaut, B., Mouchart, M., Martin, E. and Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones A comparative approach. Accident Analysis and Prevention, Issue 53, pp. 991-1004.

Fotheringham, A., Brunsdon, C. and Charlton, M., 2000. Quantitative Geography: Perspectives on Spatial Data Analysis. 1st Edition ed. SAGE Publications Ltd.

Gomes, L., Cunto, F. and Silva, A., 2017. Geographically weighted negative binomial regression applied to zonal level safety performance models. Accident Analysis and Prevention, Volume 106, pp. 254-261.

Gonzalez-Rivera, G., 2016. Forecasting for Economics and Business. Routledge.

Hadayeghi, A., Shalaby, A. and Persaud, B., 2002. Macrolevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems. Transportation Research Record: Journal of the Transportation Research Board, Volume 1840.

Hadayeghi, A., Shalaby, A. and Persaud, B., 2010. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accident Analysis and Prevention, Volume 42, pp. 676-688.

Huang, H., Abdel-Aty, M. A. and Darwiche, A., 2010. County-Level Crash Risk Analysis in Florida: Bayesian Spatial Modeling. Transportation Research Record: Journal of the Transportation Research Board, Issue 2148, pp. 27-37.

Huang, H., Song B., Xu P., Zeng Q., Lee J., and Abdel-Aty M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography, Volume 54, pp. 248-256.

Karl, K., Brunner, M. and Yamashita, E., 2006. Influence of Land Use, Population, Employment, and Economic Activity on Accidents. Transportation Research Record: Journal of the Transportation Research Board, Issue 1953, pp. 56-64.

Lesage, J., 1999. The Theory and Practice of Spatial Econometrics. Toledo: University of Toledo, Department of Economics.

Lesage, J., 2008. An Introduction to Spatial Econometrics. McCoy College of Business Administration-Department of Finance and Economics-Texas State University-San Marcos.

Li, Z., Wang W., Liu P., Bigham J., and Ragland D., 2013. Using Geographically Weighted Poisson Regression for county-level crash modeling in California. Safety Science, Volume 58, pp. 89-97.

Lovegrove, G. and Sayed, T., 2006. Macro level collision prediction models for evaluating neighborhood traffic safety. Canadian Journal of Civil Engineering, Volume 33, pp. 609-621.

MacLeod, K., Karriker-Jaffe K., Ragland D., Satariano W., Kelley-Baker T and Lacey J., 2015. Acceptance of drinking and driving and alcohol-involved driving crashes in California. Accident Analysis and Prevention, Volume 81, pp. 134-142.

Mannering, F. and Bhat, C., 2014. Analytical Methods in Accident Research: Methodological Frontier and Future Directions. Analytic Methods in Accident Research, Volume 1, pp. 1-22.

McGuire, F., 1976. Personality Factors in Highway Accidents. Human Factors, 18(5), pp. 433-442.

Miller, H., 2004. Tobler's First law and Spatial Analysis. Annals of the Association of American Geographers, 94(2), pp. 284-289.

Mitchell, A., 2005. The ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics. ESRI Press.

Naderan, A. and Shahi, J., 2010. Aggregate crash prediction models: Introducing crash generation concept. Accident Analysis and Prevention, Volume 42, pp. 339-346.

Nakaya, T., 2016. GWR4.09 User Manual. [Online] Available at: http://docplayer.net/45949326-Gwr4-09-user-manual-gwr4-windows-application-for-geographically-weighted-regression-modelling-tomoki-nakaya.html [Accessed 20 09 2017].

Nakaya, T., Fotheringham, A., Brunsdon, C. and Charlton, M., 2005. Geographically weighted Poisson regression for disease association mapping. Statistics in Medicine, Volume 24, pp. 2695-2717.

Okabe, A., Okunuki, K. and Shiode, S., 2006. SANET: A toolbox for spatial analysis on a network. Geographical Analysis, Volume 38, pp. 57-66.

Pulugurtha, S., Duddu, V. and Kotagiri, Y., 2013. Traffic analysis zone level crash estimation models based on land use characteristics. Accident Analysis and Prevention, Volume 50, pp. 678-687.

Rhee, A., Kim, J., Jee, Y. and Ulfarsson, G., 2016. Spatial regression analyses of traffic crashes in Seoul. Accident Analysis and Prevention, Volume 91, pp. 190-199.

Romano, E., Tippetts, A. and Voas, R., 2006. Language, Income, Education, and Alcohol-Related Fatal Motor Vehicle Crashes. Journal of Ethnicity and Substance Abuse, 5(2), pp. 119-137.

Romano, E., Scherer, M., Fell, J. and drivers, E., 2015. A comprehensive examination of U.S. laws enacted to reduce. Journal of Safety Research, Volume 55, pp. 213-221.

Romley, J., Cohen, D., Ringel, J. and Sturm, R., 2007. Alcohol and Environmental Justice: The Density of Liquor Stores and Bars in Urban Neighborhoods in the United States. Journal of Studies on Alcohol and Drugs, 68(1), pp. 48-55.

Dula S., Dwyer, W. and LeVerne, G., 2007. Policing the drunk driver: Measuring law enforcement involvement in reducing alcohol-impaired driving. Journal of Safety Research, 38(3), pp. 267-272.

Shariat-Mohaymany, A., Shahri, M., Mirbagheri, B. and Matkan A., 2015. Exploring Spatial Non-Stationarity and Varying Relationships between Crash Data and Related Factors Using Geographically Weighted Poisson Regression. Transactions in GIS, 19(2), pp. 321-337.

Tobler, W., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, Volume 46, pp. 234-240.

United States Census Bureau, 2017. American Fact Finder. [Online] Available at: https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml [Accessed 21 09 2017].

Wooldridge, J., 2013. Introductory Econometrics: A Modern Approach. 5th Edition ed. Michigan State University: South-Western Cengage Learning.

Xie, Z. and Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Computers, Environment and Urban Systems, 32(5), pp. 396-406.

Xie, Z. and Yan, J., 2013. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. Journal of Transport Geography, Volume 31, pp. 64-71.

Xu, P. and Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis and Prevention, Volume 75, pp. 16-25.

2.9 Appendix 2A – Steps for Variable Selection in GWPR

Step 1: prepared the data in csv format and imported into GWR4.09 then used postal code as the key and projected a coordinate system to run the Poisson (count) model by selecting standardization of the independent variables and performing a geographical variability test.

Step 2: used crash frequency as the dependent variable and an offset of 1 (assuming at least one DUI crash in each postal code over the entire five-year period).

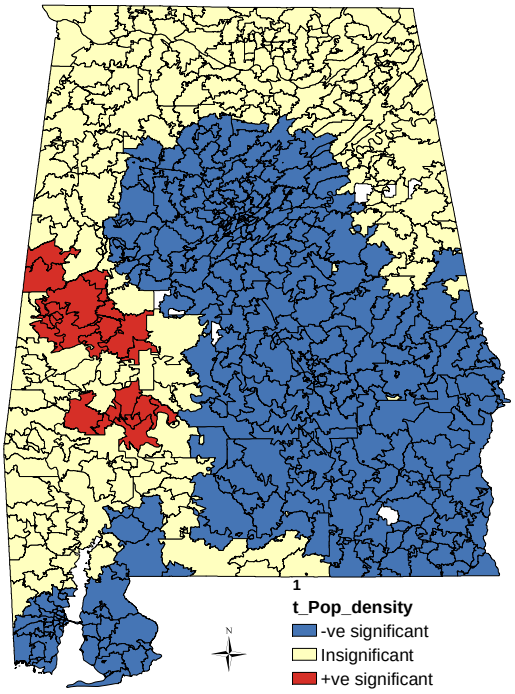
Step 3: by logical argument and reasoned forward selection, included all relevant variables as local independent variables in the initial analysis.

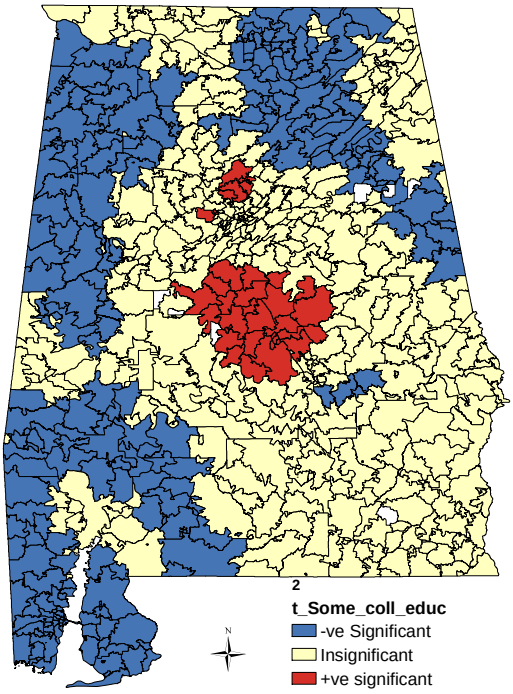
Step 4: for geographical kernel weighting, the adaptive bi-square nearest neighborhood method was used with a selection criteria of corrected Akaike Information Criterion (AIC).

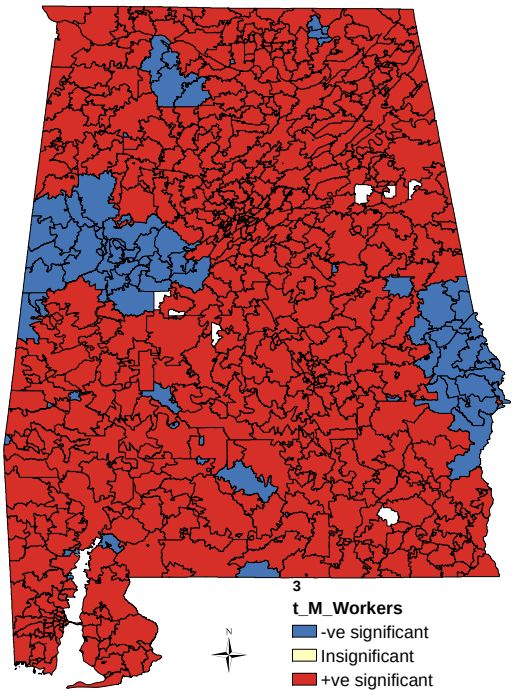
Step 5: run the model and obtained the estimates for all parameters including test of geographic variability.

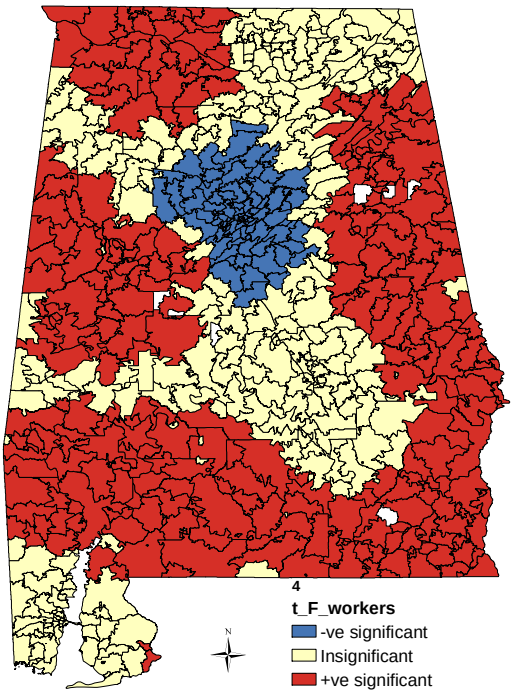
Step 6: generated maps of t-statistics for all local independent variables from the outputs. Variable that were not significant in at least 80% of the postal codes were dropped and step 1 through 6 repeated until t-statistics for all significant variables was obtained in at least 80% of the postal codes.

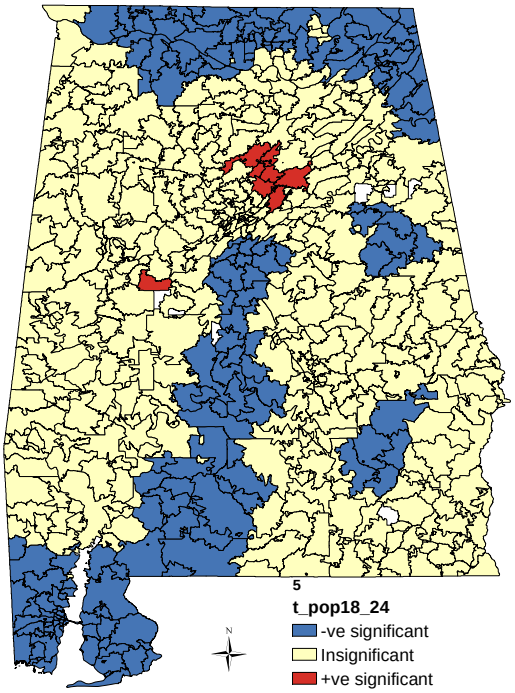
The following maps shows typical maps of t-statistics for independent variables obtained in Step 6 in one of the loops. Note that these twelve maps include some of the variables that were dropped because they were not statistically significant in at least 80% of the postal codes.

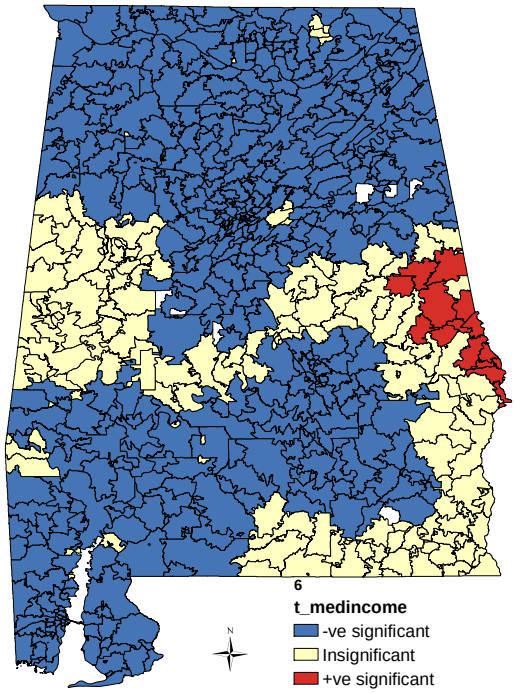

1
t_Pop_density
-ve significant
Insignificant
+ve significant

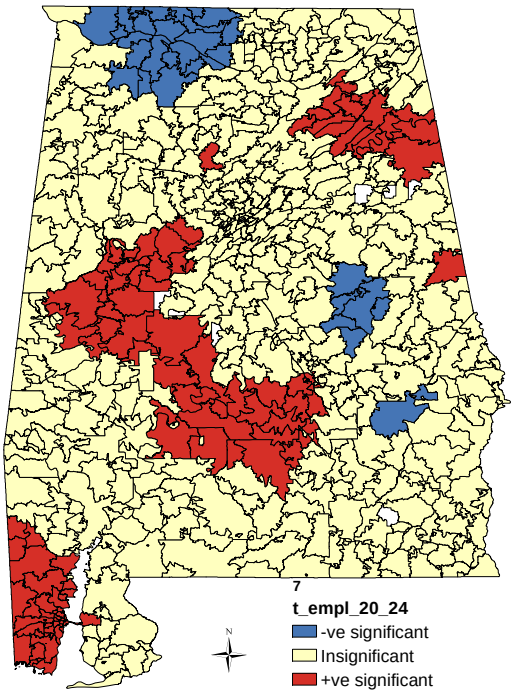

2
t_Some_coll_educ
-ve Significant
Insignificant
+ve significant

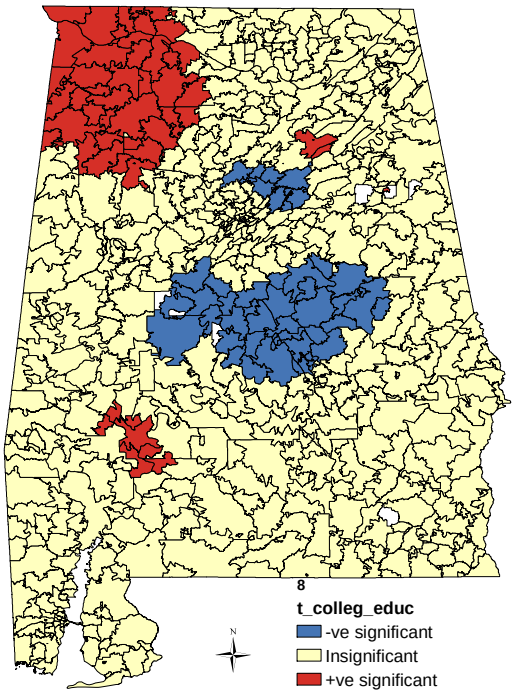

3
t_M_Workers
-ve significant
Insignificant
+ve significant

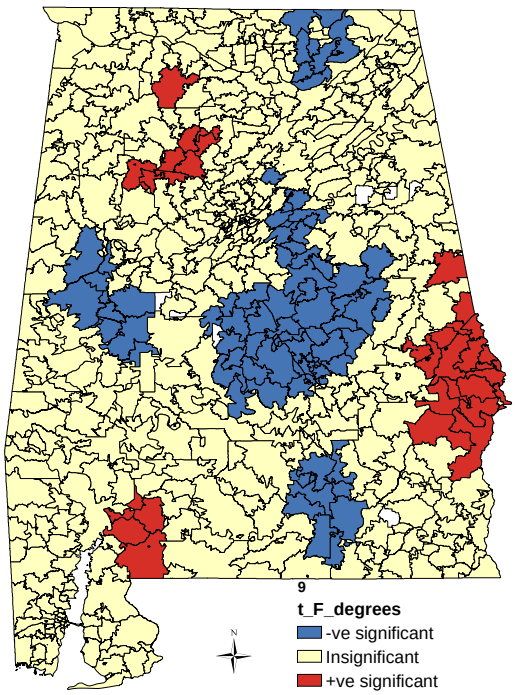

4
t_F_workers
-ve significant
Insignificant
+ve significant

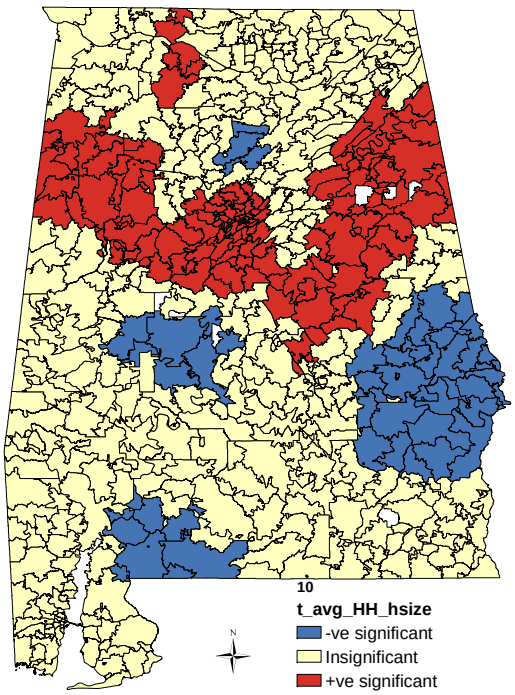

5
t_pop18_24
-ve significant
Insignificant
+ve significant

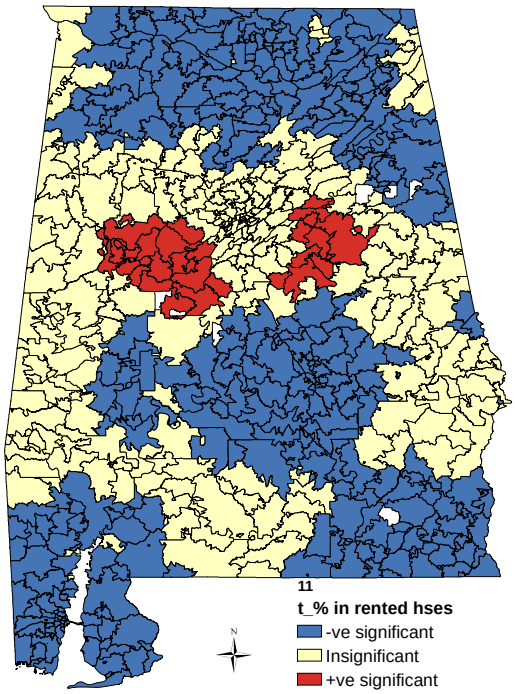

6
t_medincome
-ve significant
Insignificant
+ve significant

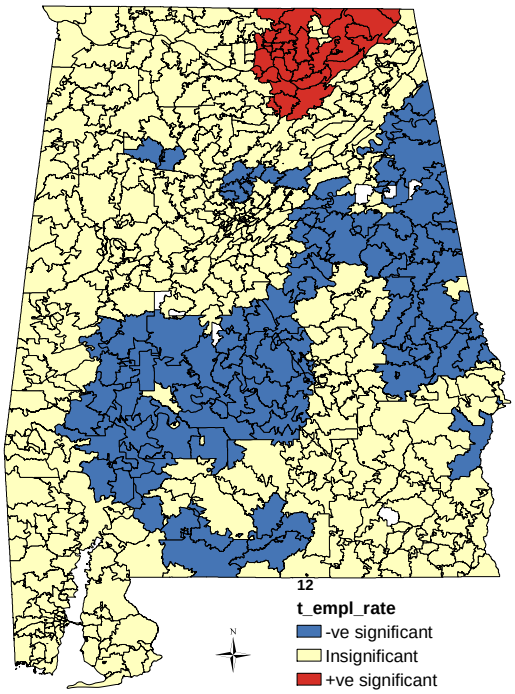

7
t_empl_20_24
-ve significant
Insignificant
+ve significant

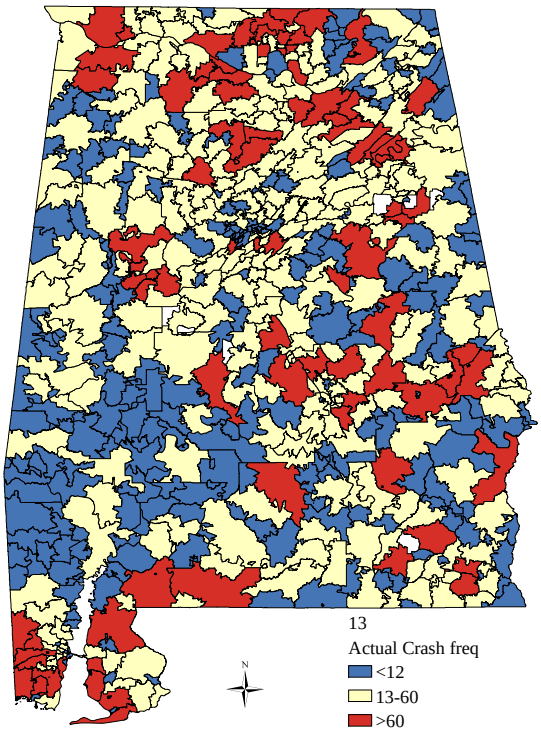

8
t_colleg_educ
-ve significant
Insignificant
+ve significant

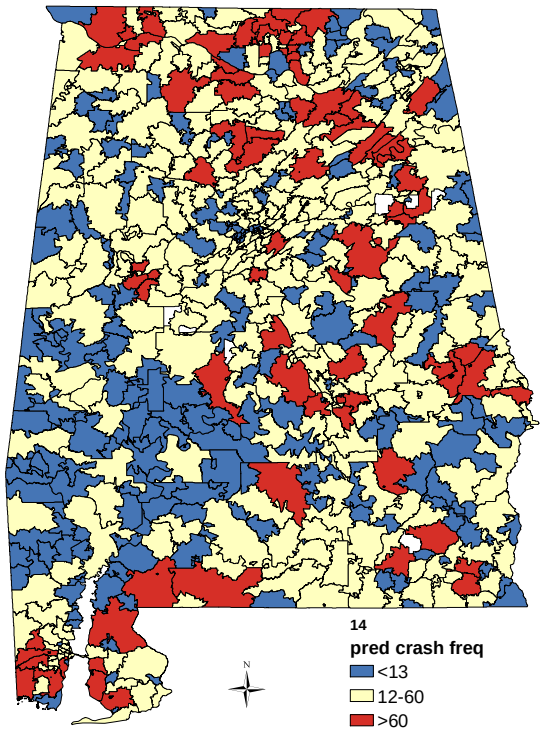

9
t_F_degrees
-ve significant
Insignificant
+ve significant


10
t_avg_HH_hsize
-ve significant
Insignificant
+ve significant


11
t_% in rented hses
-ve significant
Insignificant
+ve significant


12
t_empl_rate
-ve significant
Insignificant
+ve significant


13
Actual Crash freq
<12
13-60
>60
N


14
pred crash freq
<13
12-60
>60
N

CHAPTER 3. SPATIAL ECONOMETRIC ANALYSES OF SOCIOECONOMIC FACTORS AND TRAFFIC SAFETY – A CASE STUDY OF DUI CRASHES IN ALABAMA

3.1 Introduction

Traffic safety data and trends vary on a range of geographic scales. For example, a single crash happens at a point along a roadway. That roadway may be just inside of an area defined by local conditions such as complex terrain with substantial horizontal and vertical curve features. The local area could be in a rural region where access to emergency response is limited or where the presence of enforcement is sparse. To further complicate the issue, it has been shown that the way people drive and overall attitudes towards driving safety varies spatially (Czech, et al., 2010; Ellison, et al., 2015) and, many cases, clusters (Ozelim, et al., 2016; Adanu et al., 2017). The effects of spatial dependence and spatial heterogeneity draws from the problem of spatial aggregation, spatial externalities and spill-over effects documented throughout the econometric literature (e.g., Anselin, 1988; LeSage and Pace, 2009; Aguero-Valverde, 2013; Elhorst, 2013) and perhaps best described by a quote from the geographer Tobler (1970) that (*ceteris paribus*):

> "...everything is related to everything else, but near things are more related than distance things."

The purpose of this research is to explore the application of spatial econometric modeling techniques to identify and understand macro level socioeconomic factors influencing the occurrence of crashes.

Regardless of whether a run-off-road or rear-end crash occurs and whether it was raining or the proper traffic control equipment was operational, there are myriad underlying human-

related factors that contribute to crashes. Certainly, specific driving behaviors (e.g., aggressiveness) or activities (e.g., drunk driving) contribute to the occurrence of crashes at specific locations (i.e., event scale). Research has shown that certain human behaviors relate to overall socioeconomic (and even cultural) factors. For example, it has been shown that certain socioeconomic groups may be less likely to wear seat belts and, thus, be more susceptible to severe injuries in the occurrence of a crash. Similarly, behaviors such as drunk driving may be more acceptable (or tolerated) among certain population groups. Furthermore, it has been shown that certain common socioeconomic data (i.e., population groups) are often geographically clustered. Therefore, efforts to understand how these factors influence crashes can be conducted at the macro level (from a spatial perceptive) to determine how and where behavioral countermeasures (e.g., targeted education or selective enforcement campaigns) might be most appropriate.

This study specifically examines crashes involving drivers that were recorded as *driving under the influence* (DUI) at the time of crash. A range of macro level socioeconomic characteristics are studied at the individual postal code level using a suite of spatial econometric models. It involves the integration and analysis of large crash data sets in conjunction with socioeconomic data – all at the level of individual postal codes across the State of Alabama. It is intended that identification of socioeconomic trends affecting traffic safety can be used to inform more effective efforts and policies to target the underlying human factors causing crashes within (and across) specific regions.

3.2 Background

There have been previous efforts to study the spatial variation of factors contributing to crashes. For example, (Quddus, 2008) used spatial autocorrelation and heterogeneity analysis to

estimate a model to estimate crash counts in London. Rhee at al., (2016) used spatial regression analysis to analyze traffic crashes in Seoul by estimating a spatial lag and spatial error models. Mannering and Bhat (2014) noted that issues resulting from unobserved heterogeneity, endogeneity, as well as spatial and temporal correlations remain a methodological barrier in the statistical analysis of crash data.

Xu et al., (2017) recently used spatial econometrics to investigate spatially varying relationships between crash frequencies and related risk factors in Florida. Aguero-Valverde (2013) also applied spatial econometrics to estimate multivariate spatial models of excess crash frequency at area level in Costa Rica. There are other studies involving application of spatial econometrics to analyze crash data (e.g., Chiou, et al., 2014; Chiou and Fu, 2015; Soro, et al., 2016).

3.3 Spatial Econometrics

Spatial econometrics is a technique that combines both spatial analysis and econometric analysis. It is an extension of the traditional econometric modeling technique that considers the presence of spatial autocorrelation among neighbors. Basically, spatial econometrics is way of accounting for spatial interdependence among observations when developing statistical models to estimate relationships among dependent and independent variables (Lesage, 1999; Lesage 2008). Spatial dependence and spatial heterogeneity draws from the problem of spatial aggregation, spatial externalities and spill-over effects which is at the core of regional science and geography (Anselin, 1998). Spatial dependence simply means that observation at location i depends on other observations at locations j where $i \neq j$ and can be expressed as follows:

$$y_i = f(y_i), i = 1 \dots \dots \dots n, i \neq j \quad (3.1)$$

Basically, a "full" spatial econometric model with all the interaction effects is represented in vector form as follows (Elhorst, 2013).

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + X_i\beta + \sum_{j=1}^{n} W_{ij}\, Z\theta + \mu_i \quad (3.2)$$

$$\mu = \lambda \sum_{j=1}^{n} W_{ij}\, \mu + \varepsilon_i \quad (3.3)$$

Where:

y_i represents a continuous dependent variable corresponding to region i;

ρ represents spatial autoregressive coefficient;

λ represents the spatial autocorrelation coefficient;

θ represents the coefficient for the spatially lagged parameters for to postal code i;

μ_i represents residual from the spatial model;

X *and* Z represents independent parameters;

β *and* θ represents a vector of parameter estimates

ε_i represents error terms which are independent and irrelevantly distributed $\varepsilon_i \sim N(0, \sigma^2)$;

W_{ij} represents a matrix of spatial weights calculated as set out in section 3.4 below.

The summation terms in Equations 3.2 and 3.3 then are there to reflect the spatial variation among individual parameters as follows:

$\sum_{j=1}^{n} W_{ij}\, y_j$ represents endogenous interactions effects within the dependent variable;

$\sum_{j=1}^{n} W_{ij}\, Z$ represents exogenous interactions among independent variables; and

$\sum_{j=1}^{n} W_{ij}\, \mu$ represent any interaction among the error terms of different postal code.

By choosing which interaction effects to consider, the form of the model can change. In other words, the way in which each model treats $\rho, \lambda,$ and θ terms affects how an individual model addresses possible endogenous interactions effects, exogenous interactions effect, and/or interaction effects among error terms. The current research applied nine separate spatial

econometric models to study the relationship between macro level socioeconomic factors and DUI crashes in Alabama. The models are listed in Table 3.1 along with the treatment of the three main spatial parameters implicit to each. A full description of each of the models shown in Figure A3.1 are provided in the appendix to Chapter 3.

Table 3.1: Treatment of Spatial Lag Parameters across Models

Model		ρ	λ	θ
Spatial Durbin Error	SDEM	= 0	≠ 0	≠ 0
Spatial Durbin Moving Average	SDMA	= 0	≠ 0	≠ 0
Spatial Moving Average	SMA	= 0	≠ 0	= 0
Spatial Durbin	SDM	≠ 0	= 0	≠ 0
Spatial Durbin Autoregressive Confused	SDAC	≠ 0	≠ 0	≠ 0
Spatial Autoregressive Confused	SAC	≠ 0	≠ 0	= 0
Spatial Autoregressive Moving Average	SARMA	≠ 0	≠ 0	= 0
Spatial Durbin Autoregressive Moving Average	SDARMA	≠ 0	≠ 0	≠ 0
Spatial Autoregressive Model	SAR	≠ 0	= 0	= 0

3.4 Spatial Weights Matrix

The spatial weights matrix (W_{ij} in Equations 3.2 and 3.3) is specified using a two-step process. First the binary contiguity matrix is specified. This takes the form of a matrix containing binary outcomes for proximity of neighboring regions (postal codes in the current study). In the matrix form, the spatial ID represented by rows and columns takes a value of 1 for neighbors and value of 0 for regions that do not touch as illustrated in the sample contiguity matrix show in Figure 3.1 (Elhorst, 2013).

		Z1	Z2	Z3	Z4	Z5	…	Z_n
	Z1	0	1	0	1	0		0
	Z2	0	0	1	1	0		1
C =	Z3	0	1	0	0	1		0
	Z4	1	0	0	0	0		1
	Z5	0	0	1	0	1		0
	.							
	.							
	.							
	Z_n	0	1	0	0	1		0

Figure 3.1: Sample contiguity matrix

The diagonal values of the contiguity matrix must be zeros to avoid any region being considered as a neighbor to itself.

The final spatial weights *W*, are then created by standardizing and normalizing the contiguity matrix such that the sum of the horizontal values in the weights matrix must sum up to one. From the sample contiguity matrix above, the corresponding spatial weights matrix is illustrated in Figure 3.2.

		Z1	Z2	Z3	Z4	Z5	…	Z_n
	Z1	0.00	0.50	0.00	0.50	0.00		0.00
	Z2	0.00	0.00	0.33	0.33	0.00		0.33
W =	Z3	0.00	0.50	0.00	0.00	0.50		0.00
	Z4	0.50	0.00	0.00	0.00	0.00		0.50
	Z5	0.00	0.50	0.00	0.50	0.00		0.00
	.							
	.							
	.							
	Z_n	0.00	0.00	0.50	0.00	0.50	0.00	0.00

Figure 3.2: Spatial weights matrix formats

In this study, contiguity and spatial weights matrices were created for all the postal codes in Alabama using the process described above for use in the spatial econometric modeling process.

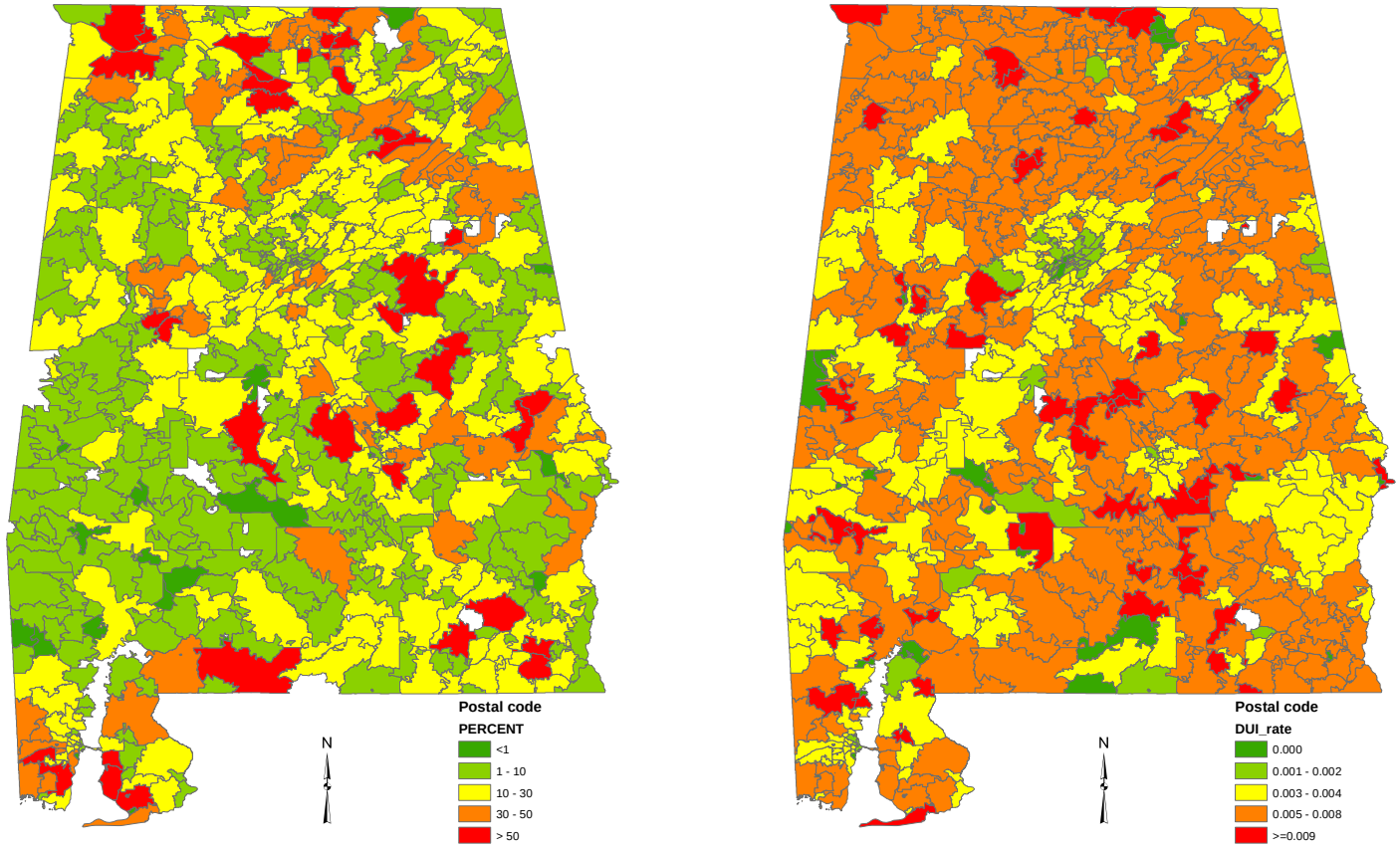
3.5 Data Description

Crash data for the State of Alabama were obtained from the Critical Analysis Reporting Environment (CARE) software developed by the Center for Advanced Public Safety (CAPS) at the University of Alabama. Each crash record contained all details related to a crash recorded by the police at the time of the crash for more than 647,000 crashes over the 2009 – 2013 study period. For the current study, the data were filtered resulting in a total of 21,818 crash records where DUI were the primary contributing circumstance. Each crash record contained the postal code of "Driver 1" indicating the driver was determined to be "at-fault" by the reporting police officer. The DUI crashes were then sorted by postal code. The mean DUI crash rate (crashes/population) per postal code was determined to be 13.9 with a standard deviation of 25.11. Socioeconomic data were obtained from the US Census Bureau (U.S. Census Bureau, 2017). Population-based crash rates were computed by dividing the DUI crash frequencies by the population of residents in each postal code. Table 3.2 summarizes the relevant crash data and 46 socioeconomic data categories assembled for the study. Figure 3.3a shows the percentage of crashes caused by drivers in a zip code that were DUI crashes and Figure 3.3b shows the DUI crash rate by population for each zip code.

Table 3.2 Variables considered in Spatial Econometric Models

Variable Description (by postal code)	Mean (Std. Dev)	Variable Description (by postal code)	Mean (Std. Dev)
Crash rates normalized by population	0.01(0.01)	Ln (males between ages 15 to 17)	1.24(0.73)
% of population between ages 15 to 17	4.08(3.20)	Ln (males who are separated)	0.55(0.85)
% of females between ages 15 to 17	3.82(4.05)	Ln (population who are separated)	0.81(0.74)
% of males between ages 15 to 44	37.76(13.02)	Ln (Crash rates normalized by population)	-2.76(0.42)
% of females between ages 15 to 44	36.49(12.71)	Ln (population who are living in rented housing)	2.90(0.94)
% of population aged above 65	16.23(10.02)	Ln (% of females who have less than high school certificate)	2.09(1.43)
% of males aged above 65	14.28(9.89)	Ln (female population)	7.38(1.67)
% of females aged above 65	18.11(11.13)	Ln (% of male with bachelor's degree or higher)	0.64(1.03)
Employment rate	47.68(13.33)	Ln (% of male who have less than high school certificate)	2.46(1.43)
% of residents living in rented housing	24.43(16.58)	Ln (population of residents with less than high school education)	2.51(1.24)
% of population living in their own housing	74.19(18.54)	Ln (population who have bachelor's degree or higher)	0.97(1.10)
% of population who are married	49.56(15.69)	Ln (unemployment rate)	2.26(0.84)
% of population who are divorced	11.70(5.37)	Median income ($10,000)	4.02(1.79)
% of population who are never married	12.69(24.75)	Ln (divorced population who are black)	1.85(1.21)
% of males who are married	51.96(16.57)	Ln (divorced population who are white)	2.24(0.82)
% of males who are divorced	11.39(6.48)	Employment rate of population who are between ages 16 to 19	2.43(1.35)
% of females who are married	48.36(16.68)	Ln (male population)	7.33(1.68)
Median income	40227(17857.29)	Ln (population between ages 18 to 24)	2.00(0.77)
Employment rate for females between ages 20 to 64	54.85(16.14)	Ln (males between ages 18 to 24)	2.01(0.85)
Employment rate for males between ages 20 to 64	64.48(19.30)	Ln (females between ages 18 to 24)	1.88(0.83)
Average household size (persons/household)	2.55(0.39)	Ln (population between ages 15 to 44)	3.53(0.61)
% of female residents with bachelor's degree or higher	6.22(11.80)	Ln (female worker force)	6.25(1.88)
% of female with some college education	41.96(26.09)	Ln (male work force)	6.40(1.85)
% of all population with only high school certificate	31.85(18.95)	Ln (male who are never married)	0.71(0.97)
Employment rate of population who are between ages 20 to 24	50.86(26.55)	Ln (females who are divorced)	2.26(0.79)
Total population	7498.73(8801.94)	Ln (females who are never married)	0.80(0.99)
Ln (% of females with only high school certificate)	2.70(1.43)	Ln (males with some college education)	2.89(1.42)
Ln (% of females who are separated)	0.88(0.84)	Ln (residents with some college education)	3.25(1.19)
Ln (males with only high school certificate)	3.03(1.35)		

Ln: Natural logarithm, %: Percentage



(a) (b)

Figure 3.3: DUI crash frequencies (a) and population-based crash rates (b) by postal code.

For analysis purposes, the DUI crash rate was a transformed into a continuous variable, CR, as shown in Equation 1 to satisfy the Gauss-Markov OLS assumptions for econometric model estimation as follows (Wooldridge, 2013). The properties of the resulting CR variable are shown in the histogram and scatter plot (by postal code) shown in Figures 3.4 and 3.5, respectively.

$$CR = Ln\{(Crash\_rate)^{0.5}\} \quad (3.4)$$

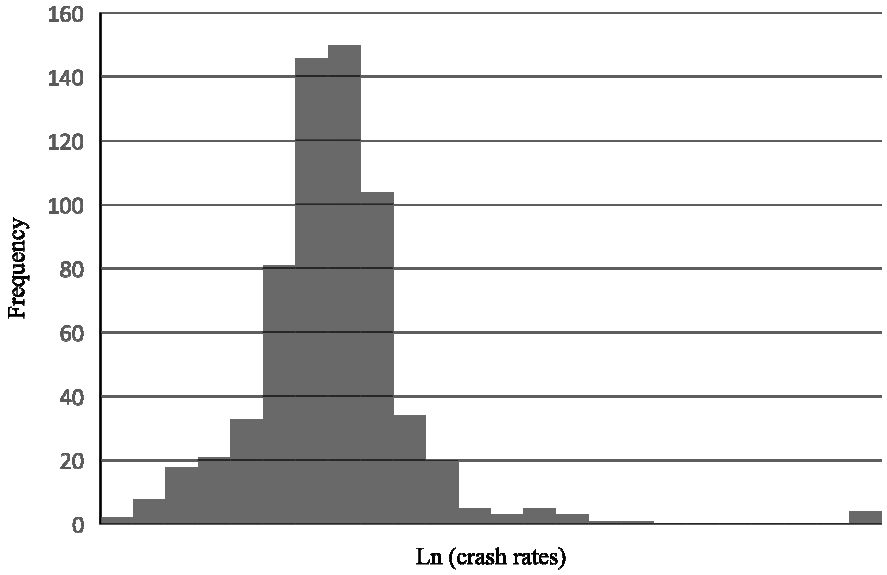


Figure 3.4: Histogram of Ln (CR) across individual postal codes

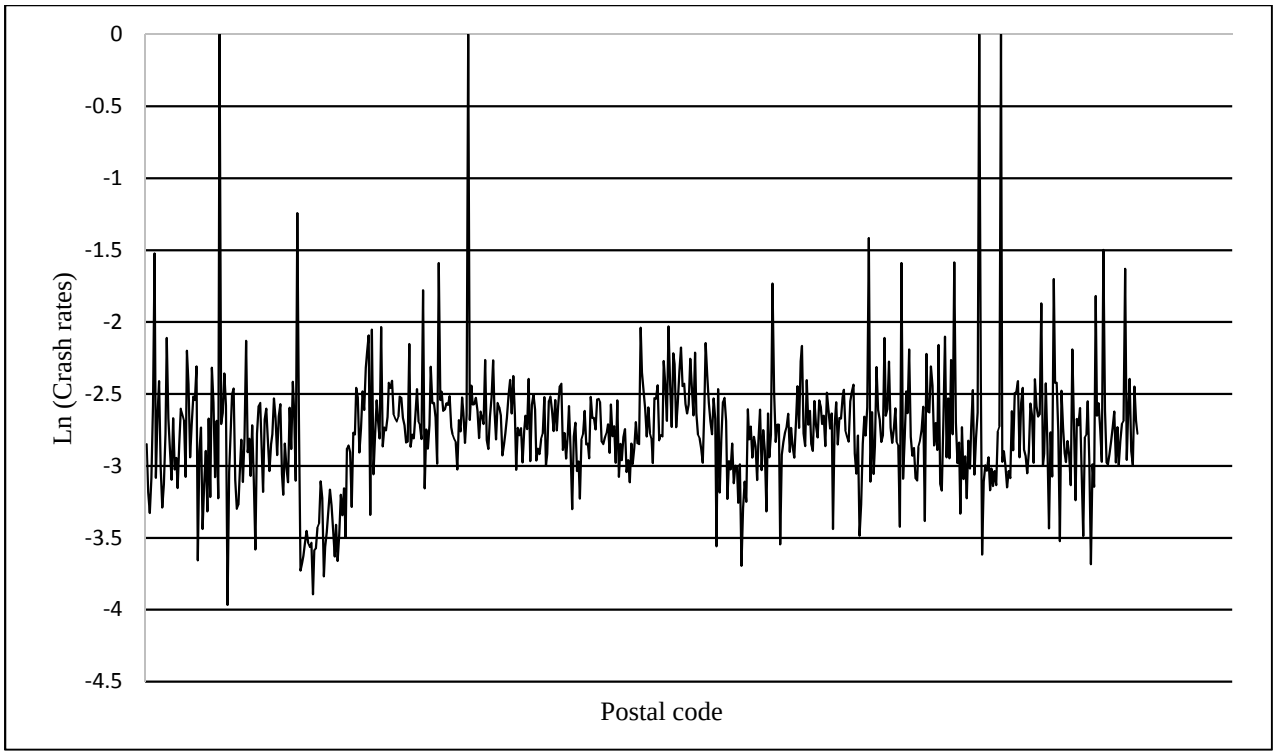


Figure 3.5: Scatter of CR across individual postal codes

Figures 3.4 and 3.5 clearly show that the transformed crash variable approximates a normal distribution. This implies it is suitable for estimation using OLS methodologies. Figure 3.5 presents an up and down movement around a central tendency line which implies that the data is stationery. As such, it depicts no specific pattern or trend which makes it appropriate for model estimation without any transformation.

3.6 Results and Discussion

Nine spatial econometric models were estimated. The models were estimated using SAS Enterprise Guide version 7.1. After developing the suite of nine spatial econometric models to estimate crash rates for individual postal codes, a total of eight independent variables were determined to be significant in addition to two variables capturing spatial lag properties. The final modeled parameters are summarized in Table 3.3.

Table 3.3: Descriptive statistics of model variables

Variable Description (by postal code)	Mean (Std. Dev.)	Variable Name
Dependent Variable		
Natural log of crash rates normalized by population in postal code	-2.757 (0.418)	CR
Employment		
Employment rate	47.678 (13.331)	EmpRate
Employment rate for ages 20 – 24	50.86 (26.546)	EmpU24
Natural log of the male work force	6.402 (1.849)	EmpMale
Family and Housing		
Percentage of residents living in rented housing	24.43 (16.581)	PRental
Average household size (persons/household)	2.547 (0.388)	HHSize
Education		
Percentage of female residents with Bachelor's degree or higher	6.215 (11.802)	EdFem
Natural log of residents with some college education.	3.246 (1.195)	PHiEd
Income		
Median income ($10,000)	4.023 (1.786)	Income
Spatial lag		
Percentage of residents living in rented housing	24.43 (16.581)	RentLag
Natural log of population of all residents with less than high school education	2.51 (1.242)	HiEdLag

The estimated results from the nine spatial econometric models are presented in Tables 3.4 – 3.6 where each table displays the results of three models. The tables include coefficients, standard errors (in parentheses) and the t-statistics for each parameter found to be significant in each model. The models are not directly comparable due to the difference in the heteroscedasticity attributed to the residuals. These heteroscedasticities are, however, taken care of in the spatial effects, spatial dependence, spatial error terms and spatial lag analysis. Nonetheless, significant parameters can be discussed together and compared one to another. It is perhaps of interest to note that the signs of the mean parameter values are the same for all significant variables which is not coincidentally consistent with other findings, for example (Castro, et al., 2012).

Table 3.4. Results for SAR, SDM and SDEM

Variable Description	***SAR***		***SDM***		***SDEM***	
	Coefficient	*t-statistics*	*Coefficient*	*t-statistics*	*Coefficient*	*t-statistics*
Intercept	-1.581 (0.098)	-16.170	-1.625 (0.094)	-17.260	-1.553 (0.090)	-17.310
Employment						
EmpRate	-		-		-	
EmpU24	-		0.002 (0.001)	3.440	0.002 (0.001)	2.770
EmpMale	-		-0.068 (0.012)	-5.610	-0.072 (0.012)	-6.200
Family and Housing						
PRental	-0.005 (0.001)	-6.150	-0.003 (.001)	-2.840	-0.002 (0.001)	-2.690
HHSize	-0.157 (0.036)	-4.400	-0.133 (0.035)	-3.810	-0.150 (0.034)	-4.480
Education						
EdFem	-0.004 (0.001)	-3.020	-0.003 (0.001)	-2.890	-0.003 (0.001)	-3.050
PHiEd	-0.122 (0.012)	-9.810	-0.104 (0.014)	-7.310	-0.101 (0.013)	-7.510
Income						
Income	-0.036 (0.008)	-4.370	-0.020 (0.009)	-2.270	-0.022 (0.009)	-2.520
Spatial Lag Effects						
RentLag	-		-0.003 (0.001)	-2.900	-0.004 (0.001)	-3.500
HiEdLag	-		0.089 (0.019)	4.690	0.058 (0.014)	4.130
rho (spatial lag coefficient)	0.029 (0.011)	2.700	0.038 (0.020)	1.920	-	
lambda (spatial error coefficient)	-				0.310 (0.042)	7.370
sigma2 (variance)	0.117 (0.007)	17.870	0.107 (0.006)	17.870	0.097 (0.005)	17.790

Standard errors are in parentheses. Level of significance is at 95%. Parameter values that were not significant at 95% are omitted.

Table 3.5: Results for SMA, SDMA and SAC

Variable Description	***SMA***		***SDMA***		***SAC***	
	Coefficient	*t-statistics*	*Coefficient*	*t-statistics*	*Coefficient*	*t-statistics*
Intercept	-1.621 (0.092)	-17.520	-1.894 (0.052)	-36.450	-1.967 (0.051)	-38.390
Employment						
EmpRate	0.004(0.001)	3.060	-		-	
EmpU24	-		0.002 (0.001)	3.480	-	
EmpMale	-0.080 (0.010)	-7.840	-0.081(0.012)	-6.960	-	
Family and Housing						
PRental	-	-	-0.002 (0.001)	-2.600	-0.004 (0.001)	-4.690
HHSize	-0.164 (0.034)	-4.770	-		-	
Education						
EdFem	-0.003 (0.001)	-3.060	-0.003 (0.001)	-2.710	-0.003 (0.001)	-2.730
PHiEd	-0.088 (0.014)	-6.510	-0.100 (0.014)	-7.220	-0.132 (0.012)	-10.810
Income						
Income	-0.020 (0.009)	-2.230	-0.027 (0.009)	-3.000	-0.046 (0.009)	-5.390
Spatial Lag Effects						
RentLag	-		-0.004 (0.001)	-3.440	-	
HiEdLag	-		0.059 (0.014)	4.190	-	
rho (spatial lag coefficient)	-		-	-	0.019 (0.010)	1.910
lambda (spatial error coefficient)	-0.330 (0.044)	-7.570	-0.261 (0.044)	-5.890	0.274 (0.043)	6.320
sigma2 (variance)	0.107 (0.006)	17.810	0.105 (0.006)	17.830	0.112 (0.006)	17.810

Standard errors are in parentheses. Level of significance is at 95%. Parameter values that were not significant at 95% are omitted

Table 3.6: Results for SDAC, SARMA and SDARMA

Variable Description	***SDAC***		***SARMA***		***SDARMA***	
	Coefficient	*t-statistics*	*Coefficient*	*t-statistics*	*Coefficient*	*t-statistics*
Intercept	-1.985 (0.052)	-38.340	-1.965 (0.051)	-38.320	-2.075 (0.048)	-42.900
Employment						
EmpRate	-		-		-	
EmpU24	-		-		-	
EmpMale	-		-		-	
Family and Housing						
PRental	-0.004 (0.001)	-4.540	-0.004 (0.001)	-4.970	-	
HHSize	-		-		-	
Education						
EdFem	-0.003 (0.001)	-2.790	-0.003 (0.001)	-2.710	-0.004 (0.001)	-3.210
PHiEd	-0.134 (0.012)	-10.960	-0.130 (0.012)	-10.560	-0.140 (0.012)	-11.430
Income						
Income	-0.045 (0.009)	-5.290	-0.045 (0.009)	-5.340	-0.037 (0.008)	-4.430
Spatial Lag Effects						
RentLag	-				-0.005 (0.001)	-4.030
HiEdLag	0.045 (0.021)	2.100			0.065 (0.021)	3.130
rho (spatial lag coefficient)	0.053 (0.019)	2.790	0.021 (0.010)	2.120	*0.035 (0.021)	1.660
lambda (spatial error coefficient)	0.250 (0.045)	5.530	-0.242 (0.045)	-5.410	-0.230 (0.045)	-5.130
sigma2 (variance)	0.112 (0.006)	17.820	0.116 (0.007)	17.840	0.116 (0.007)	17.840

Standard errors are in parentheses. Level of significance is at 95%. Parameter values that were not significant at 95% are omitted. *SDARMA parameters at the 90%.

The constant term is quite significant with a negative sign in all the models and ranges from minimum value of -2.075 to maximum value of -1.553. This implies that, *ceteris paribus*, any given (i.e., a randomly selected) driver form any given postal code is not likely to be involved in a DUI crash. The finding is consistent with previous research on adult driving populations that confirmed that drivers generally avoid drinking and driving (Shinar, Schechtman, and Compton, 2001). The following sections discuss the coefficient estimates for the significant variables and the corresponding impact on DUI crashes. The discussions are based on the broader socioeconomic groups of the parameters.

Each of the models summarized in Tables 3.4 – 3.6 contain ten significant macro level parameters classified into five main socioeconomic categories; employment, housing, education, and income. Goodness of fit statistics for all models are summarized in Table 3.7 including rankings of each model according to individual fitness measures. The results indicate that the SDEM model provides the best fit according to the conventional measures reported. The SDMA and SMA models then appear to alternate between the second and third rankings while the others maintain a consistently relative fitness ranking from SDM at fourth to SAR at ninth. Table 3.7 also shows the number of macro level parameters each model found significant out of the ten found to be significant across the suite of nine models.

Table 3.7: Goodness of Fit Statistics

Model	Log Likelihood	Rank	AIC	Rank	SBC	Rank	# Significant Parameters
SDEM	-168.87	1	361.73	1	415.25	1	7
SDMA	-182.59	2	387.18	2	436.24	3	6
SMA	-187.96	3	393.92	3	434.06	2	6
SDM	-191.99	4	407.98	4	461.50	4	7
SDAC	-210.39	5	438.78	5	478.92	6	5
SAC	-212.56	6	441.13	6	476.81	5	4
SARMA	-215.84	7	447.69	7	483.36	7	5
SDARMA	-216.14	8	450.27	8	490.41	8	4
SAR	-221.46	9	458.92	9	494.60	9	5

As noted by Nochajski and Stasiewicz (2006), DUI drivers are a heterogeneous group and cannot be defined using one model. As such, it would be insufficient to simply identify the top ranked (i.e., best fitting) model and interpret its significant parameters. Table 3.8 summarizes the role the final eight macro level parameters play in each of the models in terms of whether the individual parameter increases "+" or decreases "-" the DUI crash for a given postal code. The following sections, then, provide a discussion of each of the category of parameters that includes observations gleaned from all nine models rather than confine the discussion to just the top-ranked ones in terms of fit.

Table 3.8: Goodness of Fit Statistics

Model	EmpRate	EmpU24	EmpMale	PRental	HHSize	EdFem	PHiEd	Income
SDEM		+	-	-	-	-	-	-
SDMA		+	-	-		-	-	-
SMA	+		-		-	-	-	-
SDM		+	-	-	-	-	-	-
SDAC				-	-	-	-	-
SAC				-		-	-	-
SARMA				-	-	-	-	-
SDARMA					-	-	-	-
SAR				-	-	-	-	-

3.6.1 Employment

Table 3.8 indicates that four of the models (SDM, SDEM, SMA, and SDMA) suggest that employment-related issues influenced DUI crashes. The SMA model was the only one that showed any relationship to the overall employment rate in a postal code; the observation being that higher employment levels increase DUI crashes. The other three showed a similar relationship between employment levels among persons 20 – 24 years of age, a parameter not found significant in the SMA model. These four models all showed that an increase the male participation in the workforce decreased DUI crashes. These results are interesting in that it shows that some employment factors within a postal does affect the rate at which residents in those postal codes cause DUI crashes.

It is perhaps of interest to note that the "best fitting" model, SDEM, estimated effects for the employment of younger people while the two next best models, SDMA and SMA, differed in the some of the employment-related parameters found to be significant. All three top models, however, agreed that the increasing employment among males appears to reduce DUI crashes.

3.6.2 Family and Housing

Table 3.8 shows the top performing model found both family and housing parameters to be significant. Specifically, the results showed that seven of the models indicate a negative relationship between the percentage of rental housing in a postal code and the rate of DUI crashes caused by its residents. This is an interesting and perhaps counterintuitive finding. When viewed in terms of macro level spatial characteristics, however, it becomes more informative. Further analysis of the socioeconomic data indicates that the more populous postal codes (i.e., more urbanized) have larger percentages of rental housing. And, as noted in 3.6.5 below, the spatial lag parameter for the rental housing variable indicated significant clustering. In other words, postal codes with higher rental housing percentages exhibited lower DUI crash rates and this relationship appeared to influence DUI crash rates in neighboring postal codes. Such results seem to suggest that the rate of drivers causing DUI crashes is higher in rural areas of Alabama than in urban area – perhaps indicating a different attitude (i.e., acceptance) toward drunk driving among these communities. A different set of seven models showed that as the average household size increases the rate of DUI crashes decreases. Such a result is interesting in that households with more people (e.g., families) exhibit less propensity towards drunk driving.

3.6.3 Education

Perhaps one of the most interesting findings of the study is that all the models estimated indicate that a more educated population contributes to a lower DUI crash rate for a given postal code. Specifically, the percentage of college educated women in a postal code and the overall percentage of postal code residents with at least a high school education was shown to reduce the occurrences of DUI crashes both locally and globally.

3.6.4 Income

As with education, Table 3.8 show that all nine of the models found the average median income of a postal code to significantly reduce the DUI crash rates attributable to its residents. Clearly, higher incomes likely correlate to higher education levels. The finding however, perhaps speaks to the larger cultural issue of an overall tendency to engage in risky behaviors among high income individual (e.g., Romano et al, 2006; Factor et al, 2008).

3.6.5 Spatial Dependence and Spatial Effects

The results confirm that there is spatial autocorrelation among the dependent variable and the error terms. This is indicated by the statistically significant spatial lag coefficient and spatial error coefficient. This shows that there is spatial correlation among drivers causing DUI crashes in each postal code. As indicated, the spatial lag coefficient is statistically significant in all the models which confirms that any of the nine models is appropriate for the data. In addition, the lag coefficient is positive in all scenarios which indicates that there is a positive association among frequency of drivers who cause DUI crashes. Similarly, the spatial error coefficient is statistically significant which indicates the presence of spatial heterogeneity. In addition, variance of the error term is statistically significant in all the models which indicate that it has a non-zero variability.

There are two factors with spillover effects on DUI crash rates. Percentage of people living in rental housing and the percentage of residents with less than high school education. This observation highlights the difference between rural and urban clusters. Most clusters in urban areas live in rental housing while most clusters in rural areas would have less than high school education. As such, urban areas register fewer DUI crashes compared to rural neighborhood.

3.7 Conclusion

This study utilizes applied spatial econometrics models to estimate crash rates and identify variables that define DUI crash frequency in a postal code. It produces unbiased coefficient estimates and identifies parameters with spillover effects. It investigated the relationship between socioeconomic parameters and crash frequency per population.

Four important macro level socioeconomic factors that influence DUI crash frequency per population include employment, family and housing, education, and income. These findings are consistent with many previous studies for example (Mayhew, et al., 1981; Abdel-Aty and Abdelwahab, 2000; Hassan, et al., 2001; McKnight and McKnight, 2003; Romley, et al., 2007; Fu, 2008; Factor, et al., 2008; Peck, et al., 2008; and Romano, et al., 2015; Machado-León, et al., 2016).

In summary, the effect of each of socioeconomic factors include:

- *Employment*

  As the rate of employment increase, DUI crashes also increase. An increase in employment rate of young residents between ages 20-24 particularly, increase crash rates. Employment of more men however, reduced DUI crashes. An increase in employment rate represent an increase in the number of people who are able to buy alcoholic drinks. As a result, it leads to an increase in the number of people who are able to drink and drive.

- *Family and Housing*

  Generally, an increase in household size reduce DUI crash frequency. Semilarly, locations where most people lived in rental housing also have fewer crash rates than locations where most people live in own housing.

- *Education*

  Basically, higher education is negatively associated with DUI crash rates. Particulalry, as more women get educated, DUI crash rates decrease.

- *Income*

  Eventhough having more income increases the purchasing power of an individual, the results show that having a high income reduce the probability of being invovled in a DUI crash.

Finally, quantification of the direct and indirect effects of both global and local spillover effects can be a subject for further research. In addition, higher orders of the Autoregressive (AR) and Moving Average (MA) models can also be investigated further.

3.8 References

AAA Foundation for Traffic Safety, 2012. 2012 Traffic Safety Culture Index, Washington, DC: AAA Foundation for Traffic Safety.

Abay, K, Paleti, R. and Bhat, C., 2013. The Joint Analysis of Injury Severity of Drivers in Two-Vehicle Crashes Accommodating Seat Belt Use Endogeneity. Transport Research Part B, pp. volume 50 pp 74-89.

Abdel-Aty, M. and Abdelwahab, H., 2000. Exploring the relationship between alcohol and the driver characteristics in motor vehicle accidents. Accident Analysis and Prevention, Volume 32, pp. 473-482.

Adanu, E., Smith, R., Powell, L. and Jones, S., 2017. Multilevel analysis of the role of human factors in regional disparities in crash outcomes. Accident Analysis and Prevention, Volume 109, pp. 10-17.

Aguero-Valverde, J., 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accident Analysis and Prevention, Volume 59, pp. 365-373.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Department of Geography and Economics, University of California, Santa Barbara: Kluwer Academic Publishers.

Anselin, L., 2001. “Spatial Econometrics.” In A Companion to Theoretical Econometrics.: Oxford: Wiley-Blackwell.

Behnood, A., Roshandeh, A. and Mannering, F., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research, pp. (3-4) 56-91.

Bjorholm, S., Svenning, J., Skov, F. and Balslev, H., 2008. To what extent does Tobler's 1st law of geography apply to macroecology? A case study using American palms (Arecaceae). BMC Ecology, 8(11).

Castro, M., Paleti, R. and Bhat, C., 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, Volume 46, pp. 253-272.

Chaloupka, F. J., Saffer, H. and Grossman, M., 1993. Alcohol-Control Policies and Motor-Vehicle Fatalities. The Journal of Legal Studies, 22(1).

Chiou, Y, Chiang Fu and Chih-Wei, H., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Analytic Methods in Accident Research, Volume 2, pp. 1-11.

Chiou, Y. and Fu, C., 2015. Modeling crash frequency and severity with spatio-temporal dependence. Analytic Methods in Accident Research, 5(6), pp. 43-58.

Constantinou, E. et al., 2011. Risky and aggressive driving in young adults: Personality matters. Accident Analysis and Prevention, pp. 43 pp 1323-1331.

Czech, S., Shakeshaf, A, Byrnes, J.and Doran, C., 2010. Comparing the cost of alcohol-related traffic crashes in rural and urban environments. Accident Analysis and Prevention, Volume 42, pp. 1195-1198.

Elhorst, J., 2013. Spatial Econometrics: From Cross-Sectional Data to Spatial Panels. Berlin: Springer.

ESRI_Press, 2017. Incremental Spatial Autocorrelation. [Online] Available at: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/incremental-spatial-autocorrelation.htm [Accessed 09 Aug 2017].

Factor, R., Mahalel, D. and Yair, G., 2008. Inter-group differences in road-traffic crash involvement. Accident Analysis and Prevention, Volume 40, pp. 2000-2007.

Fu, H., 2008. Identifying repeat DUI crash factors using state crash records. Accident Analysis and Prevention, 40(6), pp. 2037-2042.

Gonzalez-Rivera, G., 2016. Forecasting for Economics and Business: Routledge.

Greenberg, M., Morral, A. and Jain, A., 2005. Drink-driving and DUI recidivists' attitudes and beliefs: a longitudinal analysis. Journal of Studies on Alcohol, 66(5), pp. 640-647.

Hassan, T., Vinod Kumar, M. and Vinod, N., 2016. Influence of demographics on risky driving behavior among powered two-wheeler riders in Kerala, India. Transportation Research Part F-Traffic Psychology and Behavior, 46(part A), pp. 24-33.

Khan, G., Xiao Qin, P. and David A. Noyce, P, 2008. Spatial Analysis of Weather Crash Patterns. Journal of Transportation Engineering, 134(5), pp. 191-202.

Kim, K. and Levine, N., 1996. Using GIS to Improve Highway Safety. Computers, Environment and Urban Systems, 20(4-5), pp. 289-302.

Lesage, J., 1999. The Theory and Practice of Spatial Econometrics. Toledo: University of Toledo, Department of Economics.

Lesage, J, 2008. An Introduction to Spatial Econometrics.: McCoy College of Business Administration-Department of Finance and Economics-Texas State University-San Marcos.

LeSage, J. and Pace, R., 2009. Introduction to Spatial Econometrics. Boca Raton: CRC Press Taylor Francis Group LLC.

Machado-León, J. et al., 2016. Socio-economic and driving experience factors affecting drivers' perceptions of traffic crash risk. Transportation Research Part F: Traffic Psychology and Behavior, Volume 37, pp. 41-51.

MacLeod, K. et al., 2015. Acceptance of drinking and driving and alcohol-involved driving crashes in California. Accident Analysis and Prevention, Volume 81, pp. 134-142.

Mannering, F. and Bhat, C., 2014. Analytical Methods in Accident Research: Methodological Frontier and Future Directions. Analytic Methods in Accident Research, Volume 1, pp. 1-22.

Mayhew, D., Warren, R., Simpson, H. and Haas, 1981. Young Driver Accidents: Magnitude and characteristics of the problem. Traffic Injury Research Foundation of Canada.

McGuire, F, 1976. Personality Factors in Highway Accidents. Human Factors, 18(5), pp. 433-442.

McKnight, A. and McKnight, A., 2003. Young novice drivers: Careless or clueless? Accident Analysis and Prevention, 35(6), pp. 921-926.

Mehta, G. et al., 2014. Analyzing crash frequency and severity data using novel techniques. A Dissertation. [Online] Available at: http://acumen.lib.ua.edu/content/u0015/0000001/0001738/u0015_0000001_0001738.pdf [Accessed 26 10 2017].

Miller, H, 2004. Tobler's First law and Spatial Analysis. Annals of the Association of American Geographers, 94(2), pp. 284-289.

Nochajski, T. and Stasiewicz, P., 2006. Relapse to driving under the influence (DUI): a review. Clinical Psychology Review, 26(2), pp. 179-195.

Ozelim, L. et al., 2016. Factors in differential safety performance across different income levels. Rio de Janeiro, Brazil, 17-19 May 2016, 17th International Conference Road Safety on Five Continents (RS5C 2016).

Peck, R., Gebers, M., B. Voas, R. and Romano, E., 2008. The relationship between blood alcohol concentration (BAC), age, and crash risk. Journal of Safety Research, 39(3), pp. 311-319.

Quddus, M., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accident Analysis and Prevention, Volume 40, pp. 1486-1497.

Rhee, k, Kim, J., Jee, Y. and Ulfarsson, G, 2016. Spatial regression analyses of traffic crashes in Seoul. Accident Analysis and Prevention, Volume 91, pp. 190-199.

Romano, E., Tippetts, A. and Voas, R, 2006. Language, income, education, and alcohol-related fatal motor vehicle crashes. Journal of Ethnicity and Substance Abuse, 5(2), pp. 119-137.

Romano, E., Scherer, M, Fell, J. and Taylor, E, 2015. A comprehensive examination of U.S. laws enacted to reduce. alcohol related crashes among underage drivers Journal of Safety Research, Volume 55, pp. 213-221.

Romley, J, Cohen, D., Ringel, J. and Sturm, R., 2007. Alcohol and environmental justice: The density of liquor stores and bars in urban neighborhoods in the United States. Journal of Studies on Alcohol and Drugs, 68(1), pp. 48-55.

Dula, S., Dwyer W. and Laverne, G., 2007. Policing the drunk driver: Measuring law enforcement involvement in reducing alcohol-impaired driving. Journal of Safety Research, 38(3), pp. 267-272.

SAS Institute Inc., 2016. SAS/ETS® 14.2 User's Guide. The Spatial reg Procedure. [Online] Available at: https://support.sas.com/documentation/onlinedoc/ets/142/spatialreg.pdf [Accessed Monday September 2017].

Shawky, M., Al-Badi, Y. and Al-Ghafli, A., 2017. Relationship between Socio-Demographic of Drivers and Traffic Violations and Crashes Involvements. Barcelona, Spain, Proceedings of the 2nd World Congress on Civil, Structural, and Environmental Engineering.

Shinar, D., Schechtman, E. and Compton, R., 2001. Self-reports of safe driving behaviors in relationship to sex, age, education and income in the US adult driving population. Accident Analysis and Prevention, Volume 33, pp. 111-116.

Soro, W., Zhou, Y. and Wayoro, D., 2016. Crash rates analysis in China using a spatial panel model. IATSS Research.

Stephens, A., Bishop, C., Liu, S. and Fitzharris, M., 2017. Alcohol consumption patterns and attitudes toward drink-drive behaviors and road safety enforcement strategies. Accident Analysis and Prevention, Volume 98, pp. 241-251.

Tobler, W., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, Volume 46, pp. 234-240.

U.S. Department of Transportation. Federal Highway Administration, Dec 2008. Office of Operations. [Online] Available at: http://www.ops.fhwa.dot.gov/publications/fhwahop09005/quick_clear_laws.pdf [Accessed 25 July 2016].

United States Census Bureau, 2017. American Fact Finder. [Online] Available at: https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml [Accessed 21 09 2017].

Wooldridge, J, 2013. Introductory Econometrics: A Modern Approach. 5th Edition ed. Michigan State University: South-Western Cengage Learning.

World Health Organization (WHO), 2014. Global Status Report on Alcohol and Health, Geneva: WHO.

World Health Organization, 2015. Global Status Report on Road Safety, Geneva, Switzerland.: World Health Organization.

Xu, P. and Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis and Prevention, Volume 75, pp. 16-25.

Xu, P., Huang, H., Dong, N. and Wonga, S., 2017. Revisiting crash spatial heterogeneity: A Bayesian spatially varying coefficients approach. Accident Analysis and Prevention, Volume 98, pp. 330-337.

3.9 Appendix 3 – Taxonomy of Spatial Econometric Models

The following section presents details of ten spatial econometric models followed by Figure A3.1 that depicts how the various models are related. The first one represents a basic linear model and, although foundational to the others, it is not used in the current study because the coefficient estimates would not be BLUE due to violation of the Gauss-Markov assumptions of No-serial correlation, multicollinearity and deterministic parameters. The remaining nine models represent the ones used in the previously presented analyses.

3.9.1 Linear Model

The linear model can be described in vector form as:

$$y_i = X_i\beta_i + \varepsilon_i, i = 1 \dots \dots \dots n$$

where $\varepsilon_i \sim N(0, \sigma^2)$, X_i is a $p \times 1$ vector that denotes the values of p parameters at unit i and β is a $px1$ parameter vector.

3.9.2 Spatial Autoregressive (SAR) Model

The SAR model accounts for spatial dependence in the dependent variable- the endogenous interaction effect. In matrix notation, the first-order spatial autoregressive model is expressed in vector form as (Anselin, 1988; Anselin, 2001):

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + X_i\beta + \varepsilon_i$$

And the standard estimator is the SAR is the maximum likelihood estimator (MLE) given as (Anselin, 2001) and (SAS Institute Inc., 2016):

$$L = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{(Ay - X\beta)'(Ay - X\beta)}{2\sigma^2} + ln|A|$$

Where ρ is spatial autoregressive coefficient, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1,2 \ldots n\ zip\ codes, X_i$ is a $p \times 1$ vector that denotes the values of p parameters at postal code unit i and β is a $px1$ parameter vector. W_{ij} is the $(i,j)th$ element of the weights matrix W subject to $W_{ij} \neq 0$. Which also denotes the spatial weight between postal code $i\ and\ j$. $A = (I_n - \rho W)$ with I_n being an $n \times n$ identity matrix. $|A|$ denotes the determinant of A.

3.9.3 Spatial Durbin Model (SDM)

The SDM accounts for exogenous and endogenous interaction effects. It considers spatially lagged dependent variables as well as spatially lagged explanatory variables. It shows if inclusion of lagged parameters is warranted or not warranted in the model specification (Anselin, 1988). It allows the model to include factors in the region of observation plus the same factors averaged over the neighboring regions (Lesage, 2008). In general, SDM model is given in vector form as (LeSage and Pace, 2009):

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + X_i\beta + W_{ij}Z\theta + \varepsilon_i$$

By letting $\tilde{X} = [X\ \ WZ]$ and $\tilde{\beta} = (\beta'\ \ \theta')$ the SDM model can be re-written as:

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + \tilde{X}_i\tilde{\beta} + \varepsilon_i$$

Further, the standard estimator is the SDM is the maximum likelihood estimator (MLE) given as (Anselin, 2001):

$$L = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{(Ay - \tilde{X}\tilde{\beta})'(Ay - \tilde{X}\tilde{\beta})}{2\sigma^2} + ln|A|$$

Where Z is a $q \times 1$ vector denoting the values of q parameters measured in postal code i. θ is the coefficient for the spatially lagged parameters with respect to postal code i. The descriptions of the other variables are the same as explained above (SAS Institute Inc., 2016).

3.9.4 Spatial Durbin Error Model (SDEM)

SDEM takes care of exogenous interaction effect and spatial dependence among the error terms. The SDEM is described using two stage formulation as (LeSage and Pace, 2009):

$$y_i = X_i\beta + W_{ij}Z\theta + \mu_i$$

$$\mu = \lambda W_{ij}\mu + \varepsilon_i$$

By letting $\tilde{X} = [X \ \ WZ]$ and $\tilde{\beta} = (\beta' \ \ \theta')$ the SDM model can be re-written as:

$$y_i = \tilde{X}_i\tilde{\beta} + \beta^{-1}\varepsilon_i$$

Where $B = (I_n - \lambda W)$ with I_n being an $n \times n$ identity matrix. The log likelihood estimation for SDEM is given as follows (Anselin, 2001).

$$L = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{[B(y - \tilde{X}\tilde{\beta})]'[B(y - \tilde{X}\tilde{\beta})]}{2\sigma^2} + ln|B|$$

$|B|$ denotes the determinant of B. And the remaining variables are as already discussed above.

3.9.5 Spatial Moving Average Model (SMA)

Like any other moving average model, the SMA takes care of local autocorrelation by accounting for spatial dependence among the error terms. The vector formulation is given as (LeSage and Pace, 2009):

$$y_i = X_i\beta + \mu$$

$$\mu = (I_n - \lambda W_{ij})\varepsilon_i$$

I_n denotes a vector of n constant values. The log likelihood estimation is given as follows (Anselin, 2001).

$$L = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{[B^{-1}(y - X\beta)]'[B^{-1}(y - X\beta)]}{2\sigma^2} - ln|B|$$

All variables are as already described above. Note that, in the SEM, the residual is a function of the error terms- it depends on the error terms of the regression. While, in the SMA the error term is a function of the residual of the error term function- it takes the form of a traditional moving average function. See (Gonzalez-Rivera, 2016) for more details regarding moving average models.

3.9.6 Spatial Durbin Moving Average Model (SDMA)

The SDMA model compares to SMA model but it also accounts for spatially lagged explanatory variables- that is exogenous interaction effects. The general vector formulation for the SDMA is given as (SAS Institute Inc., 2016):

$$y_i = X_i\beta + W_{ij}Z\theta + (I_n - \lambda W)\,\varepsilon_i$$

By letting $\tilde{X} = [X \;\; WZ]$ and $\tilde{\beta} = (\beta' \;\; \theta')$ the SDMA model can be re-written as:

$$y_i = \tilde{X}_\iota\tilde{\beta} + B\varepsilon_i$$

And the log-likelihood function for model estimation is given as (Anselin, 2001):

$$L = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{\left[B^{-1}\left(y - \tilde{X}\tilde{\beta}\right)\right]'\left[B^{-1}\left(y - \tilde{X}\tilde{\beta}\right)\right]}{2\sigma^2} - ln|B|$$

All variables are as already described above.

3.9.7 Spatial Autoregressive Confused Model (SAC)

Spatial Autoregressive Confused (SAC) model takes care of spatial dependence in the dependent variable as well spatial dependence in the error term (LeSage and Pace, 2009). In a nut shell, it is a combination of Spatial Autoregressive (SAR) model and Spatial Error Model (SEM). The SAC cab be implemented with a single spatial weights matrix in which $W_1 = W_2 = W$ (Lesage, 2008). This approach handles the influence that the parameter would have within the postal code and between the postal codes. The formulation can be presented in vector form as follows (SAS Institute Inc., 2016):

$$y_i = \rho W_1 y_j + X_i \beta + \varepsilon_i$$

$$\mu = \lambda W_2 \mu + \varepsilon_i$$

The model is estimated using the log-likelihood function given as follows (SAS Institute Inc., 2016):

$$L = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{[B(Ay - X\beta)]'[B(Ay - X\beta)]}{2\sigma^2} + ln|A| + ln|B|$$

Where $B = (I_n - \lambda W_2)$ and $A = (I_n - \rho W_1)$ and all variables are as already described above.

3.9.8 Spatial Durbin Autoregressive Confused Model (SDAC)

Spatial Durbin Autoregressive Confused (SDAC) model is an extension of the SAC model which accounts for the exogenous effect. That is, it takes care of the dependence among observations, dependence among error terms and the influence that the neighboring parameters have on the dependent variable, all at the same time in one model (SAS Institute Inc., 2016). SDAC is an extension of SDM which nests the SAC model in it and is given in vector form as follows (LeSage and Pace, 2009):

$$y_i = \rho W_1 y_j + X_i\beta + W_1 Z\theta + \mu_i$$

$$\mu = (I_n - \lambda W_2)^{-1}\varepsilon_i$$

Like SDM, by letting $\tilde{X} = [X \;\; W_1 Z]$ and $\tilde{\beta} = (\beta' \;\; \theta')'$ the SDAC model can be re-re-written as:

$$y_i = \rho W_1 y_j + \tilde{X}_i\tilde{\beta} + (I_n - \lambda W_2)^{-1}\varepsilon_i$$

The model is estimated by log-likelihood estimation method given as (SAS Institute Inc., 2016):

$$L = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{\left[B\left(Ay - \tilde{X}\tilde{\beta}\right)\right]'\left[B\left(Ay - \tilde{X}\tilde{\beta}\right)\right]}{2\sigma^2} + ln|A| + ln|B|$$

All the used variables are as already explained above.

3.9.9 Spatial Autoregressive Moving Average Model (SARMA)

SARMA is like SMA model but it accounts for spatial dependence among error terms as well as spatial dependence among the dependent variable. It is basically a combination of the SMA model and the SAR model (LeSage and Pace, 2009). This can be represented in vector form as follows (Lesage, 2008):

$$y_i = \rho W_1 y_j + X_i\beta + \mu_i$$

$$\mu = (I_n - \lambda W_2)\varepsilon_i$$

The log-likelihood function for the SARMA is given as (SAS Institute Inc., 2016):

$$L = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{[B^{-1}(Ay - X\beta)]'[B^{-1}(Ay - X\beta)]}{2\sigma^2} + ln|A| - ln|B|$$

All the used variables are as already explained above.

3.9.10 Spatial Durbin Autoregressive Moving Average Model (SDARMA)

SDARMA accounts for spatial dependence among error terms as well as spatial dependence among the dependent variable and considers the influence of the exogenous parameters, the influence of the parameters within the postal code and between the postal codes. The spatial weights W_1 and W_2 can be identical and equal to W or can be different. In this study, $W_1 = W_2$. The vector formulation is given as shown below (SAS Institute Inc., 2016):

$$y_i = \rho W_1 y_j + X_i\beta + W_1 Z\theta + \mu_i$$

$$\mu = (I_n - \lambda W_2)\varepsilon_i$$

The log-likelihood function for the SDARMA is given as (SAS Institute Inc., 2016):

$$= -\frac{n}{2} ln(2\pi\sigma^2) - \frac{\left[B^{-1}\left(Ay - \tilde{X}\tilde{\beta}\right)\right]'\left[B^{-1}\left(Ay - \tilde{X}\tilde{\beta}\right)\right]}{2\sigma^2} + ln|A| - ln|B|$$

All the used variables are as already explained above.

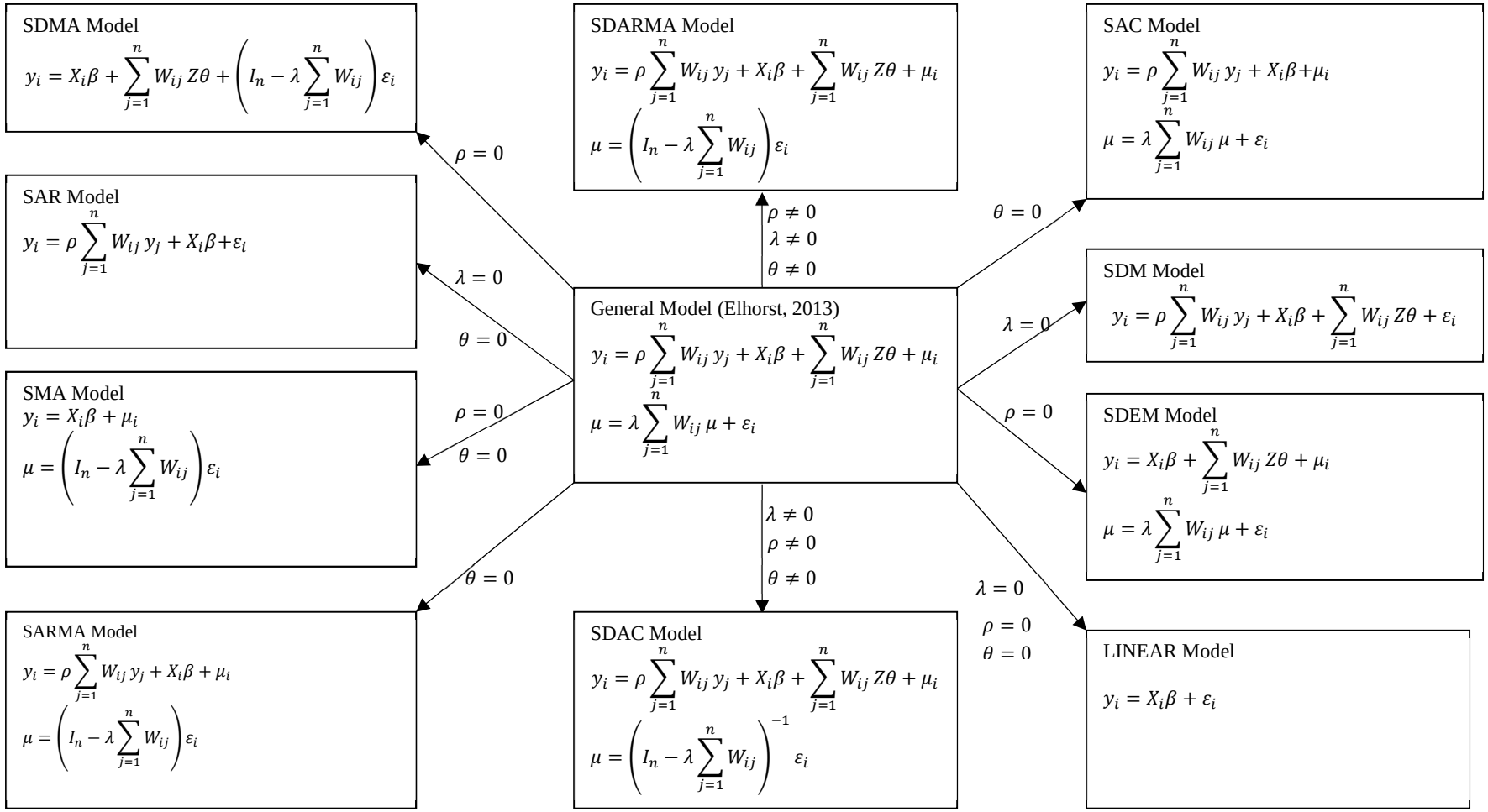


Figure A3.1: Relationships among different spatial econometric model forms

CHAPTER 4. ANALYSIS OF THE IMPACTS OF ROAD CRASHES ON FREEWAY CONGESTION AND MOBILITY – A CASE STUDY OF INTERSTATE 65 IN ALABAMA

4.1 Introduction

This research presents a comprehensive analysis of the relationship between traffic accidents and congestion. Traffic accidents are not only a leading cause of death (World Health Organization, 2015) but also a major cause of nonrecurring congestion (Wanga, et al., 2013). Crashes occurring on congested urban freeways result in significant delays to commuters, freight traffic, and all travelers (Jones, et al., 1991). The Federal Highway Administration (U.S. Department of Transportation. Federal Highway Administration, Dec 2008) estimated that Traffic Incident Management efforts in USA are credited with reducing annual delay by 129.9 million hours with an associated cost savings of $2.5 billion. Despite the efforts, traffic incidents are not only frequent but also life threatening to motorists and responders, particularly regarding secondary crashes (Wang, et al., 2016). Additionally, it is estimated that between 2003 and 2007, some 70 emergency responders (e.g., police) and 54 maintenance personnel have lost their lives after being struck by vehicles along the highway (U. S. Department of Transportation. Federal Highways Administration, 2008). At the same time, the Towing and Recovery Association of America (TRAA) reported a loss of more than 100 towing operators in the line of service.

Effective incident management requires an understanding of incident patterns, frequency and duration (Giuliano, 1989). The benefit of reducing incident duration by one minute is estimated as $1,320 per incident (Adler, et al., 2013). The subject of incident congestion duration has been a topic of study by previous researchers (e.g., Dickerson, et al., 2000; Quddus, et al.,

2010; Hojati, et al., 2013; Hojati, et al., 2014). Sullivan et al (2013) used crowd sourced vehicle speed data and accident report records to measure nonrecurring congestion along Interstate 65 in Birmingham area, Alabama. Factors affecting the crash timeline from occurrence to clearance, taking into consideration police notification, police arrival time up to the time police leaves crash scene have also been studied (e.g., Drakopoulos, et al., 2001). In addition, Hojati et al (2013) listed various methods used for estimating incident duration models such as: linear regression analysis, non-parametric regression methods and tree-style classification models, support vector regression, conditional probability analyses, probabilistic distribution analyses, time sequential methods, discrete choice models, Bayesian classifier, fuzzy logic models, and artificial neural networks.

Previous research has shown that incident prediction models can improve management of nonrecurring congestion (Garib, et al., 1997). Incident duration can be predicted by several factors including: the number of lanes affected, number of vehicles involved, truck involvement, time of day, police response time, and weather condition (Garib, et al., 1997). A study by Giuliano (1989) showed that major of explanatory factors for incident duration includes incident type, time of day, truck involvement, and lane closures. An accurate prediction of incident duration can help traffic operators implement appropriate mitigation measures (Pereira, et al., 2013). On the same note, Mekker et al (2015) analyzed three years of crash data and crowd sourced probe data to classify crashes associated with certain kinds of queue conditions and traffic flow conditions on interstates in Indiana.

The general objective of this research is to develop and test a methodology to identify factors that influence the magnitude (i.e., severity) of congestion attributed to crash events at a network level using statistical modeling techniques. After presenting the conceptual development

of the proposed methodology, its potential use is illustrated through a case study application of interstate crash data in Alabama.

4.2 Conceptual Approach

To analyze the traffic congestion attributable to crash events, this study defines a new measure called speed differential mile hours (SDMH) defined by Equation 4.1.

$$SDMH = \left\{\frac{\sum_{m_i}^{m_f}\sum_{t_i}^{t_f}(FFS_i - Speed_{m_t})}{60}\right\} * miles \quad (4.1)$$

Where:

m_i and m_f indicate start and end milepost,

t_i and t_f are the initial and end time stamps, and

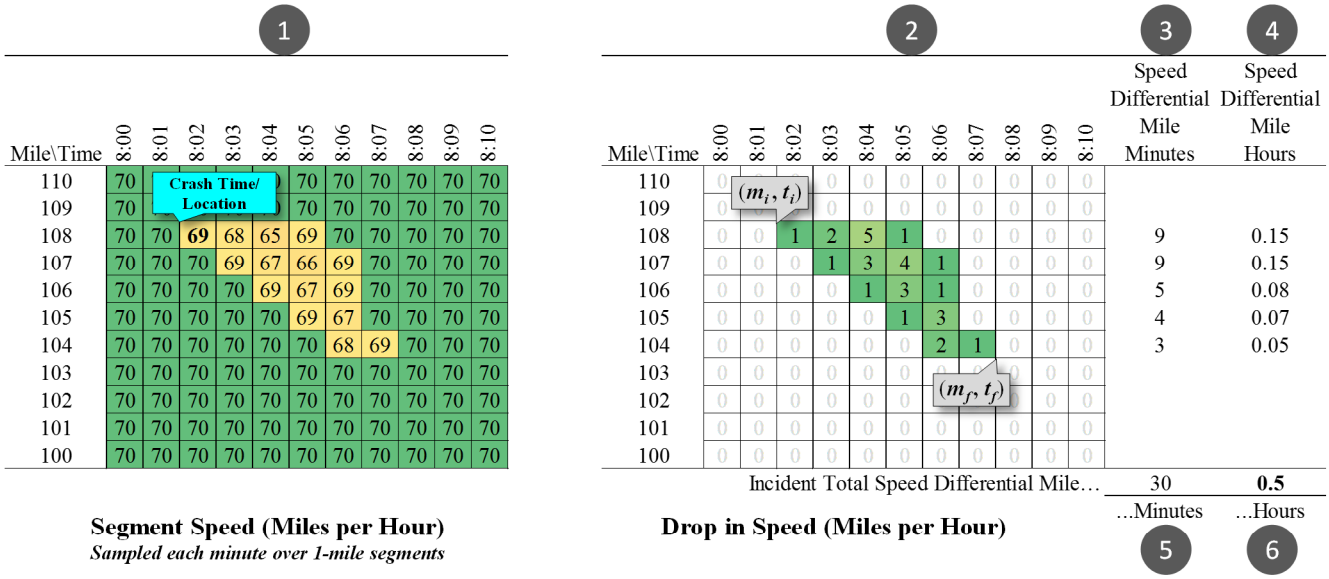
FFS_i is the free flow speed at road segment i.

The SDMH is calculated by analyzing a segment upstream of a crash location following a crash event using the six-step process described below:

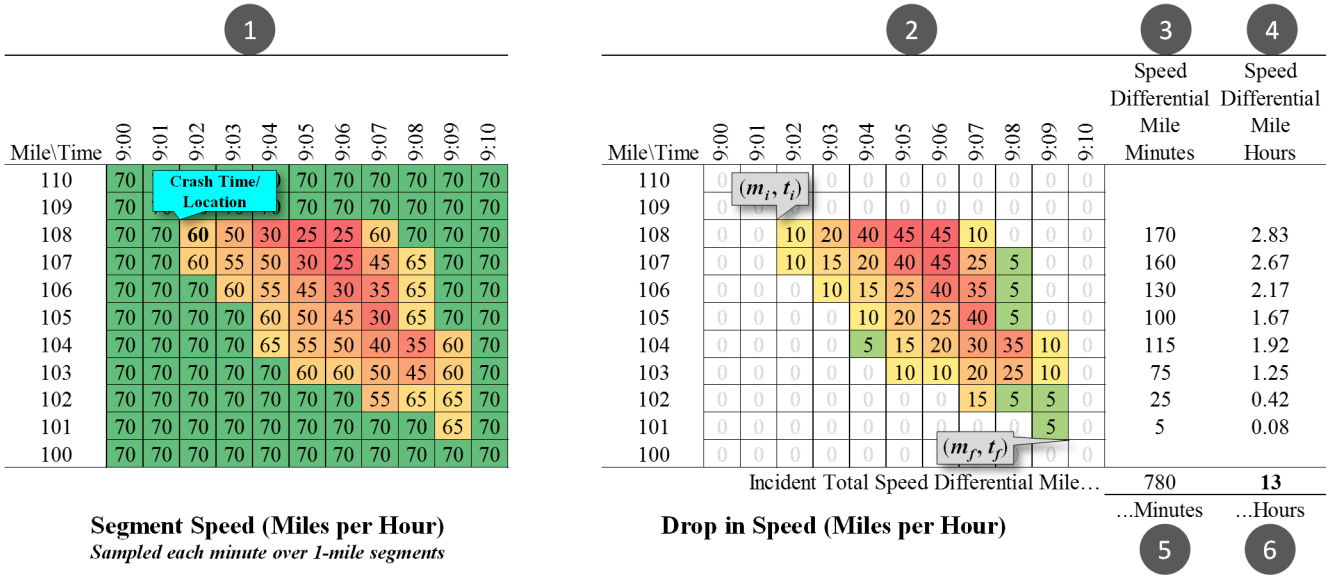
- Step 1 – traffic speed data is collected (from vehicle probe data) for the segment where the crash event occurs.
- Step 2 – estimate reduction in free flow speed (per mile per minute) attributed to the crash.
- Step 3 – sum reduction in free flow speed for each minute and each segment length.
- Step 4 – divide the reduction in free flow speed per mile per minute by 60 to obtain the reduction in speed per mile per hour per segment.

- Step 5 – sum up the values obtained in Step 3 which gives the speed differential mile minute.
- Step 6 – divide value obtained in Step 5 by 60 to obtain the SDMH.

Figure 4.1 illustrates two sample applications of this this process. Figure 4.1a shows how the process is followed step-by-step to calculate an SDMH of 0.5 and provides a visual image of the time-space domain. Figure 4.1b, then illustrates how a more intense speed reduction and longer upstream impact results in a larger SDMH being calculated – in this case and SDMH of 13.



(a) NB Crash at 8:02AM (SDMH = 0.5)



(b) NB Crash at 9:02AM (SDMH = 13)

Figure 4.1: Estimation of the SDMH for a crash event

4.3 Data Description

An SDMH was calculated for each 4,814 crash events recorded on Interstate 65 (I-65) in Alabama during 2014. The crash data was obtained from the Critical Analysis Reporting Environment (CARE) developed by the Center for Advanced Public Safety (CAPS) at the University of Alabama. The crash data comprised individual crash records containing all details recorded by the police at the time of the crash including when the crash occurred, when the police arrived, and the exact location on the roadway. Crowd sourced speed data was obtained from probe vehicle based service INRIX for the entire length of I-65 through the State of Alabama. The speed data was recorded as average speeds on a second-by-second basis in 1-mile increments. Weather data for 2014 was obtained from the national weather service website to identify the exact weather conditions present during and immediately after the crash. All the three categories of data were integrated by time and location to develop a complete picture of conditions affecting each crash event and its resulting level of traffic congestion.

Figure 4.2 illustrates how the SDMH is calculated and interpreted using a space-time diagram developed from actual data. Specifically, it shows three crash-congestion events. The time and location of the crash is marked with a white diamond marker, the police response time as a black line, and the time the police arrive on site as a gray diamond.

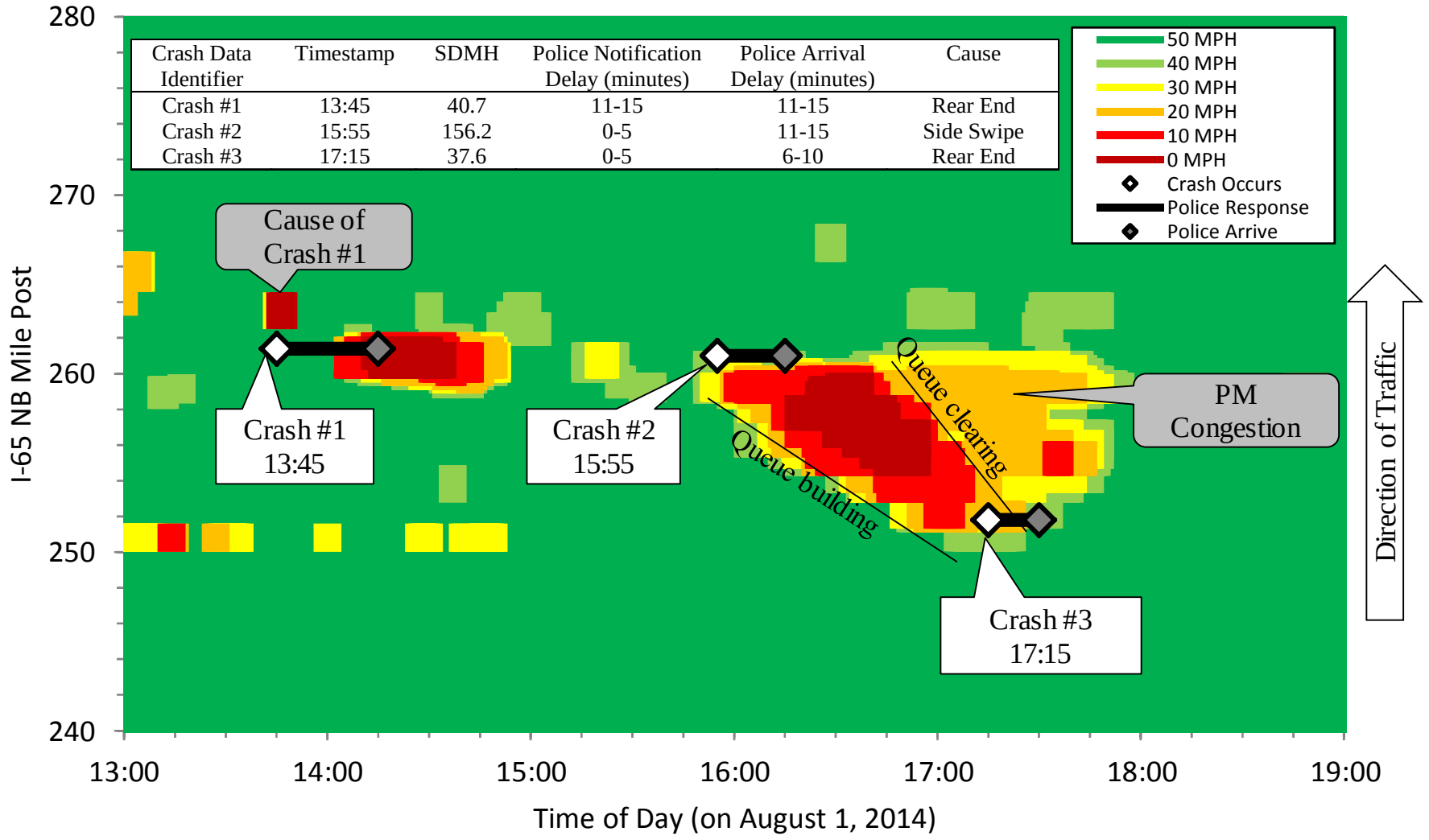


Figure 4.2: Example of a Time space diagram for a crash congestion event

Figure 4.2 shows that Crash #1 occurs within the milepost 262 traffic message channel (TMC) segment. The police response time is about 10-minutes and traffic remained stable at speeds greater than 50 mph until the police arrived. Once police arrived, traffic had to shift for the move-over law and congestion began to build. The SDMH for this crash is 40.7.

Crash #2 occurred at 15:55PM as a primary sideswipe crash (note the inset table of Figure 4.2). Based on the data, crash report, and visual depiction of queueing in Figure 4.2, traffic congestion quickly built up behind the disabled vehicles before police arrived. The queue built over about an hour to a length of about 10-miles. The SDMH for this crash is 156.2 which appropriately suggest a much greater impact or severity of congestion than crash #1.

Just as the queue from crash #2 was about to clear, crash #3 occurred as a secondary crash near the back of the crash #2 queue. Based on the data and crash report for crash #3, the vehicle was moved to the side of the road as no additional queue formed upstream of the crash location at mile post 251.8. However, the final callout on Figure 4.2 identifies residual congestion during the PM peak period near the interchange ramps. The SDMH for this crash is 37.6, which is about the same as crash #1. The congestion region for crash #2 and crash #3 is very consistent with traditional shockwave theory (May, 1990).

4.4 Methodology

The relationships among the characteristics of each interstate crash and the ensuing congestion (as measured by SDMH) were explored using statistical analysis. The following sections present the development of a dependent variable from the calculated SDMH attributable to an individual crash and the statistical technique used to estimate its relationship to crash-related parameters.

4.4.1 SDMH Variable Development

The SDMH was calculated for each of the 4,814 crashes in 2014 on Interstate 65, the SDMH was measured for each event. The SDMH was bounded to 5-miles upstream of the crash location and to a 1-hour duration after the time of the crash. These boundaries were used to help normalize the sampling region and eliminate other phenomena that can contribute to congestion other than the crash alone. The alternative to a fixed-boundary approach is using an algorithm to delineate the boundary and capture the "entire" crash, but these algorithms can be challenging under real conditions. For example, a crash occurring during peak periods with normally-anticipated recurring congestion will have part of the congestion attributed to the crash while part of it will be attributed to the peak time recurring congestion (refer to crash #2 in Figure 4.2). In another scenario, a crash occurring when there is inclement weather can be challenging to delineate the reduction in speed due to a crash versus the wide-spread drop in speed due to weather over the entire region. Therefore, to fairly estimate the impact of a crash on congestion, a fixed-boundary approach was used in this work. The SDMH approach in this paper is advantageous in that observing segments with no reduction in speed would not contribute to error in the SDMH measurement (refer to Figure 4.1 where the "0" drop in speed does not increase the overall SDMH score).

Finally, to avoid the need for a precise continuous metric, four discrete categories of congestion severities based on the SDMH scores were established to include: No congestion, Low congestion, Medium congestion, or High congestion as summarized in Table 4.1. Vehicle probe data has been used in the past to quantify and discretely categorize incident-related non-recurrent congestion on key interstate facilities in a similar fashion (Sullivan, et al., 2013).

Table 4.1: Congestion severities and SDMH band widths

Congestion severity	Fall in free flow speed (SDMH) Range (mph)	Percentage represented
No congestion	0	26%
Low congestion	1-10	27%
Medium congestion	10-50	25%
High congestion	>50	22%

4.4.2 Statistical Analysis

This study uses a mixed logit model to characterize crash congestion because it can address the limitations of multinomial logit by allowing for heterogeneous effects (unobserved heterogeneities) and correlation in unobserved factors as shown by Savoleinen, et al., (2011). The approach is also used because of the ability to provide a reasonable level of accuracy as shown by Anastasopoulos and Mannering (2011). Mixed logit assumes that choice probability is a mixture of logits with specific mixing distribution (Revelt and Train, 1997). McFadden and Train (2000) showed that it provides a flexible and computationally practical approach to discrete response analysis. Many researchers have used mixed logit models to study road safety (Train, 1998; Gkritza and Mannering, 2008; Milton, et al., 2008; Anastasopoulos and Mannering, 2009; Moore, et al., 2011; Ukkusuri, et al., 2011; Mehta and Lou, 2013; Islam and Jones, 2014; Islam, et al., 2014; Islam and Hossain, 2015). In general, mixed logit is widely used and accepted. Other models which take into consideration unobserved heterogeneity such as random parameter negative binomial, have also been estimated to increase the understanding of road safety (Chin and Quddus, 2003; Venkataraman, et al., 2011; Anastasopoulos, et al., 2012; Venkataraman, et al., 2013).

Discrete categorization of congestion has been used in past studies based on different thresholds. For example, Sullivan et al., (2013) used crowd sourced vehicle probe data and accident reports to measure nonrecurring congestion along Interstate 65 in small urban area in Birmingham Alabama by categorizing congestion severities as a discrete outcome. Given the above discrete outcomes, the probability of a crash having any congestion severity is estimated. The strength of random parameter models derives from the ability to allow parameter values to vary across population (Washington, et al., 2011). The development of mixed logit modeling approach follows the work of McFadden and Train (2000) which takes the form of the following probability outcome.

$$T_{in} = \beta_i X_{in} + \varepsilon_{in} \quad (4.2)$$

Where:

$T_{in}$is the probability of crash n having congestion severity i,

β_i is a vector of estimable parameter for severity i which may vary across population,

X_{in} is an independent variable and

ε_{in} is a disturbance term.

The above equation takes the form of a standard Multinomial form shown below.

$$P_n(i) = \frac{e^{(\beta_i X_{in})}}{\sum_{\forall I} e^{(\beta_I X_{In})}} \quad (4.3)$$

Where:

P_n is the probability of crash n having congestion severity i.

For a mixed logit model, the outcome probabilities for the above equation therefore becomes.

$$P_n^m(i) = \int_x P_n(i) f(\beta|\varphi)\, d\beta \quad (4.4)$$

Where:

$f(\beta|\varphi)$is the density function of β with φ being the vector of parameter of that density function (mean and variance).

As such, the mixed logit function can be expressed as follows (Washington et al, 2011; Anastasopoulos and Mannering, 2011).

$$P_n^m(i) = \int_x \frac{e^{(\beta_i X_{in})}}{\sum_{\forall I} e^{(\beta_I X_{In})}} f(\beta|\varphi)\, d\beta \quad (4.5)$$

The parameters are approximated using Halton draws method by randomly drawing values of β from $f(\beta|\varphi)$ and using the values to estimate a simple multinomial logit probability. Bhat (2003) showed that 200 Halton draws is sufficient to estimate the parameters. The simulated log-likelihood function is given as follows (Washington et al., 2011).

$$LL = \sum_{n=1}^{N} \sum_{i=1}^{I} \delta_{in} LN[P_n^m(i)] \quad (4.6)$$

Where:

N is the total number of crashes,

I is the total number of severities (in this case four) and

δ_{in} is defined as 1 if the observed discrete outcome for crash n is i and zero otherwise.

This study uses a normal distribution form of the density function and the simulated likelihood function is maximized.

Finally, marginal effects (for continuous covariates) are estimated sing the following equation (Mannering, 2009; Washington et al., 2011).

$$E_{X_{ki}}^{P(i)} = [1 - P(i)]\beta_{ki} X_{ki}. \quad (4.7)$$

Where:

$P(i)$ is the probability of having congestion severity (i).

β is the parameter estimate for variable X under category k in congestion severity i.

This is the effect of a 1% change in variable X on the probability of congestion severity i.

The marginal effects for indicator variables (where 0 represent a NO and 1 represent a YES) is also given as follows (Mannering, 2009; Washington et al., 2011).

$$E_{X_{ki}}^{P(i)} = \left[\frac{EXP[\Delta(\beta_i X_i)]\sum_{\forall I} EXP(\beta_{kI} X_{kI})}{EXP[\Delta(\beta_i X_i)]\sum_{\forall I_n} EXP(\beta_{kI} X_{kI}) + \sum_{\forall I \neq I_n} EXP(\beta_{kI} X_{kI})} - 1\right] * 100 \quad (4.8)$$

Where the variable descriptions are the same as for marginal effects estimates for continuous covariates.

4.5 Results

Variable selection was done using a guided forward selection process to test the hypothesis of each additional variable. Variables and transformations of variables (such as creating binary indicators and variable interactions) were added and the model adjusted accordingly. The SDMH was used as the dependent variable which represented the congestion severities as presented in Table 4.1. The analysis was done in NLOGIT4.0 software (NLOGIT4.0, 2011) and the results are presented in Table 4.2. It shows coefficient estimates and the t-statistics including the mean and standard deviation of random parameters. All the parameters included are statistically significant (statistically different from 0 at 95% level of confidence). Normal distribution provided the best fit for all the random parameters.

Table 4.2: Model Parameter Estimates

Parameter	Coefficient	Std error	Prob	t-statistics	95% Confidence Interval	
No Congestion						
AADT	-0.024	0.004	0.000	-6.000	0.032	-0.016
Dry road surface	1.183	0.197	0.000	6.000	.796	1.570
Urban area	-1.158	0.220	0.000	-5.260	1.589	-0.727
Weekend	0.717	0.169	0.000	4.240	.386	1.049
Random parameters						
Heavy trucks (Standard deviation)	-0.160	0.035	0.000	-4.580	0.229	-0.092
	(0.186)	(0.034)	(0.000)	(5.530)	(0.120)	(0.252)
Low congestion severity						
Intercept	-2.575	0.569	0.000	-4.520	3.691	-1.459
Early morning	3.118	0.475	0.000	6.570	.188	4.048
Misty weather	0.881	0.262	0.001	3.370	.369	1.394
Medium congestion severity						
Intercept	-3.997	0.575	0.000	-6.950	5.125	-2.869
Rain	6.107	0.940	0.000	6.500	.265	7.949
AADT	0.012	0.001	0.000	9.020	.009	0.014
Warning signs	2.499	0.635	0.000	3.930	.253	3.744
Vehicle towed	0.591	0.092	0.000	6.410	.410	0.772
DUI	0.687	0.243	0.005	2.830	.212	1.163
Early morning	3.765	0.478	0.000	7.880	.828	4.702
Random parameters						
Winter season (Standard deviation)	0.164	0.220	0.456	0.750	0.267	0.594
	(1.882)	(0.628)	(0.003)	(3.000)	(0.652)	(3.113)
Bridge (Standard deviation)	-0.491	0.633	0.438	-0.780	1.732	0.750
	(2.784)	(1.391)	(0.045)	(2.000)	(0.058)	(5.510)
High congestion severity						
Intercept	-6.496	0.645	0.000	-10.070	7.760	-5.233
AADT	0.016	0.002	0.000	10.520	.013	0.020
Speed	1.046	0.222	0.000	4.720	.612	1.481
Work zone workers	1.462	0.262	0.000	5.570	.947	1.976
Snow	4.235	0.427	0.000	9.910	.397	5.073
Morning peak time	1.009	0.133	0.000	7.590	.748	1.269
Afternoon peak time	1.343	0.121	0.000	11.060	1.105	1.581
Physical barrier present in median	0.659	0.124	0.000	5.340	.417	0.902
Random parameters						
Fatal Crash (Standard deviation)	0.227	0.158	0.151	1.440	0.083	0.538
	(1.646)	(0.322)	(0.000)	(5.120)	(1.015)	(2.276)
Number of Observations	4694					
Log-likelihood at convergence	-5723.436					
McFadden Pseudo R-squared	0.120					
R-squared	0.1205					

Standard deviation of random parameters is in parentheses. Std error: Standard error. Prob: Probability.

Marginal effects calculated from the analysis show the relationship between the dependent variable and the parameters. Table 4.3 is a summary of the estimated marginal effects for each significant variable.

Table 4.3: Marginal effects

		No congestion	Low	Medium	High
Heavy trucks	No congestion	-0.260	-	-	-
Road surface	No congestion	0.400	-	-	-
Urban area	No congestion	-0.426	-	-	-
Weekend	No congestion	0.071	-	-	-
Misty weather	Low congestion	-	0.008	-	-
Rain	Medium congestion	-	-	0.021	-
Warning signs	Medium congestion	-	-	0.003	-
Vehicle towed	Medium congestion	-	-	0.151	-
DUI	Medium congestion	-	-	0.009	-
Early morning	Low congestion	-	0.088	-	-
	Medium congestion	-	-	0.069	-
Winter season	Medium congestion	-	-	0.098	-
Bridge	Medium congestion	-	-	0.019	-
AADT	No congestion	-0.996	-	-	-
	Medium congestion	-	-	0.486	-
	High congestion	-	-	-	0.700
Speed	High congestion	-	-	-	0.594
Work zone workers	High congestion	-	-	-	0.020
Snow	High congestion	-	-	-	0.025
High severity crash	High congestion	-	-	-	0.370
Morning peak time	High congestion	-	-	-	0.091
Afternoon peak time	High congestion	-	-	-	0.160
Median present	High congestion	-	-	-	0.334

4.6 Discussion

The marginal effects are presented in Table 4.3. This is the effect of a 1% change in the variable, or change of category given a binary variable, on the probability of having any congestion severity. The following section presents the impact of each significant variable on congestion severity.

4.6.1 No Congestion

As presented in Table 4.1, *No congestion* is characterized by an SDMH of 0 - in which, a crash on an interstate does not results in a change in free flow speed. From the marginal effects presented in Table 4.3, a 1% increase in AADT reduces the probability of having *No congestion*

by about 99%. In other words, adding traffic to the interstate increases the chance of a crash causing congestion as would be expected. Crashes occurring during the weekend are more likely to cause *No congestion*. Table 4.1 suggests that crashes occurring over the weekend a 7% more likely to result in *No congestion* as it would reasonable be anticipated that overall volumes are lower and travel times/routes more elastic than weekdays (e.g., commuting periods). Crashes occurring in urban sections of I-65 were shown to reduce the probability of *No congestion* by approximately 43%. A crash occurring on a dry road surface condition increased the probability of having *No congestion* by about 40%. This finding agrees with other studies on relationship between speed and road surface condition (e.g., Cao, et al., 2016).

The results indicate that heavy trucks have a random effect on congestion. About 80.5% of the time, a 1% increase in heavy truck volume reduced the probability of having no congestion by 26.1% (the procedure for the marginal effect estimates from random parameters is included in Appendix 4B). Crashes involving trucks may be more likely to be severe and, perhaps more importantly, can increase the chance of blockage of one or more lanes (Grenzeback, et al., 1990). On the other hand, heavy trucks increase the probability of having no congestion by 19.5%.

4.6.2 Low Congestion

Low congestion was defined in Table 4.1 as exhibiting an SDMH value from 1 – 10. There are only two variables that significantly characterizing low congestion severity. The first one is early morning crashes which occur between 1:00 am and 5:00 am. The probability that the congestion will be low during early morning is higher by about 8.8%. In addition, misty weather conditions also appear to slightly (0.1%) increase the chance low congestion.

4.6.3 Medium Congestion

Eight factors significantly contribute to *Medium congestion* (SDMH of 11-50). Crashes involving driving under the influence (DUI) showed a slightly higher chance of resulting in medium congestion. This result is interesting when viewed considering literature noting that DUI crashes often involve injuries (e.g., Xie, et al., 2012) and that injury crashes typically increase the likelihood of congestion (Hojati, et al., 2013). A crash occurring during rainy conditions increases the probability of having a medium congestion by 2.1% and is consistent with previous findings (e.g., Andrey, 2010). In addition, a crash that involves a vehicle being towed increases the probability of a medium congestion by about 15%. This agrees with previous research findings for example (Hojati, et al., 2013) which indicated that vehicle towing increases overall incident clearance time. Crashes outside the warning sign zone were only slightly more likely to cause medium-severity congestion by about 0.3%. Early morning (1:00 am to 5:00 am) crashes were also expected to cause medium congestion. The probability of having a medium congestion increases by about 6.9% during these hours. On the other hand, a 1% increase in AADT (in 1000 vehicles) increases the probability of having a medium congestion by 48.6%. Finally, Crashes which occur at a bridge section - an overpass or underpass, have a random effect on medium congestion. For about 43% of the time, crashes occurring at a bridge section increase the probability of a medium congestion severity by about 1.9%. While for 57% of the time, crashes at a bridge site are less likely to have medium congestion. The estimates of all marginal effects are included in Appendix 4B. Finally, the winter season (December, January, and February) exerts a random effect on medium congestion. The probability of the congestion being medium during the winter increased by about 10% for more than half (53.5%) of the time.

4.6.4 High Congestion

Traffic volume (i.e., AADT) exerted the largest influence on whether a crash resulted in high congestion (SDMH $\geq$ 50). A 1% increase in AADT (or additional 1000 vehicles) increased the probability of high congestion by about 70%. In addition, a crash occurring when there is snow increased the chance of high congestion by 2.5% - it should be noted that the crash data analyzed was for only one year (2014) and there were very few snow events in that year in Alabama. The presence of workers increases the probability of having a high congestion by about 2%. In addition, crashes caused by drivers whose estimated speed was higher than 60 mph increased the probability of a high congestion severity by 59.4% as higher speeds would be expected to more likely result in more severe injuries which would increase incident clearance times. Presence of a physical barrier/median separating opposing traffic increases the probability of having a high congestion severity by approximately 33.4% as has been shown by other (e.g., Tay and Churchill, 2007). Moreover, a crash occurring during either morning and afternoon peak time 6am to 8am and 4pm to 6pm) increases the probability of having high congestion by about 9% and 16%, respectively. This is intuitive due to relatively higher traffic volumes during these hours. Finally, fatal crashes have a random effect on high congestion. For 55.5% of the time, a fatal crash increases the probability of having high congestion by about 37%. On the other hand, fatal crashes decreased the probability of having high congestion 44.5% of the time.

4.7 Conclusions

This research provides insight into the relationship between the characteristics of crashes occurring on interstate facilities and the congestion resulting from such events. It is intended that a better understanding of this relationship can help inform traffic and incident management strategies to more efficiently deploy resources (e.g., emergency responders, service

patrol vehicles). An improved understanding of the relationship between crashes and congestion can help lead to mobility improvements and important safety enhancements by reducing the propensity for secondary crashes.

4.8 References

Adler, M, Ommeren, J. and Rietveld, P., 2013. Road congestion and incident duration, The Netherlands: VU University Amsterdam.

Alabama Department of Transport (ALDOT), 2015. Alabama Speed Management Manual. [Online] Available at: http://www.dot.state.al.us/dsweb/div_ted/Traffic_SOS/pdf/Alabama%20Speed%20Management%20Manual%20October%202015.pdf [Accessed 26 10 2017].

Anastasopoulos, P., Mannering, F, Shankar, V. and Haddock, J, 2012. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis and Prevention, September, Volume 45, p. 628– 633.

Anastasopoulos, P. and Mannering, F, 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention, 41(1)

Andrey, J., 2010. Long-term trends in weather-related crash risks. Journal of Transport Geography, Volume 18, pp. 247-258.

Cao, L., Thakali, L., Fu, L. and Donaher, G., 2016. Effect of Weather and Road Surface Conditions on Traffic Speed of Rural Highways. Washington DC, pp. 13-0779.

Chin, H. and Quddus, M, 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention, November 35(2), p. 253–259.

Dickerson, A., Peirson, J. and Vickerman, R., 2000. Road Accidents and Traffic Flows: An Econometric Investigation. Economica, Volume 67, p. 101–21.

Drakopoulos, A., Shrestha, M. and Örnek, E., 2001. Freeway Crash Timeline Characteristics and Uses. Transportation Research Record: Journal of Transport Research Board, 1748(01), pp. 132-143.

Garib, A., Radwan, A. and Al-Deek, H., 1997. Estimating Magnitude and Duration of Incident Delays. Journal of Transportation Engineering, 123(6), pp. 459-466.

Giuliano, G., 1989. Incident characteristics, frequency, and duration on a high volume urban freeway. Transportation Research Part A: General, 23(5), pp. 387-396.

Gkritza, K. and Mannering, F, 2008. Mixed logit analysis of safety-belt use in single- and multi-occupant vehicles. Accident Analysis and Prevention, July 40(2), pp. 443-451.

Grenzeback, L., Reilly, W, Roberts, P. and Stowers, J, 1990. Urban Freeway Gridlock Study: Decreasing the Effects of Large Trucks on Peak-Period Urban Freeway Congestion. Transportation Research Record, Volume 1256, pp. 16-27.

Hojati, A, Ferreira, L., Washington, S. and Charles, P., 2013. Hazard based models for freeway traffic incident duration. Accident Analysis and Prevention, December, Volume 52, p. 171– 181.

Hojati, A et al., 2014. Modelling total duration of traffic incidents including incident detection and recovery time. Accident Analysis and Prevention, June, Volume 71, p. 296–305.

Islam, S., Jones, S. and Dye, D., 2014. Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. Accident Analysis and Prevention, March 67(2), p. 148–158.

Islam, S. and Hossain, A, 2015. Comparative Analysis of Injury Severity Resulting from Pedestrian–Motor Vehicle and Bicycle–Motor Vehicle Crashes on Roadways in Alabama. Transportation Research Record: Journal of the Transportation Research Board, No. 2514(09), p. 79–87.

Islam, S. and Jones, S., 2014. Pedestrian at-fault crashes on rural and urban roadways in Alabama. Accident Analysis and Prevention, August 72(1), p. 267–276.

Jones, B., Janssen, L. and Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention, 23(4), pp. 239-255.

May, A, 1990. Traffic flow fundamentals. Prentice Hall.

Mehta, G. and Lou, Y., 2013. Modeling school bus seat belt usage: Nested and mixed logit approaches. Accident Analysis and Prevention, October 51(1), p. 56– 67.

Mekker, M, Remias, S, McNamara, M. and Bullock, D, 2016. Characterizing Interstate Crash Rates Based on Traffic Congestion Using Probe Vehicle Data. Washington DC pp. Paper No. 16-1194.

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

Moore, D, Savolainen, P. and Farzaneh, M., 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis and Prevention, 43(3), p. 621–630.

NLOGIT4.0, 2011. Computer software, NY: Econometric Software.

Pereira, F, Rodrigues, F. and Ben-Akiva, M., 2013. Text analysis in incident duration prediction. Transportation Research Part C, 37(Part C), pp. 177-192.

Quddus, M, Wang, C. and Ison, S, 2010. Road Traffic Congestion and Crash Severity: Econometric Analysis Using Ordered Response Models. Journal of Transportation Engineering © ASCE, May 136(5), p. 424–435.

Revelt, D. and Train, K., 1997. Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. The Review of Economics and Statistics, October 80(4), pp. 647-657.

Sullivan, A, Sisiopiku, V. and Kallem, B, 2013. Measuring Non-Recurring Congestion in Small to Medium Sized Urban Areas, Birmingham.

Tay, R. and Churchill, A., 2007. Effect of Different Median Barriers on Traffic Speed. Canadian Journal of Transportation, 1(1), pp. 56-66.

Train, K, 1998. Recreation Demand Models with Taste Differences over People. Land Economics, May 74(2), pp. 230-239.

U.S. Department of Transportation. Federal Highway Administration, Dec 2008. Office of Operations. [Online] Available at: http://www.ops.fhwa.dot.gov/publications/fhwahop09005/quick_clear_laws.pdf [Accessed 25 July 2016].

Ukkusuri, S., Hasan, S. and Aziz, H, 2011. Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. Transportation Research Record: Journal of the Transportation Research Board, Volume No. 2237, p. 98–106.

Venkataraman, et al., 2011. Model of Relationship Between Interstate Crash Occurrence and Geometrics Exploratory Insights from Random Parameter Negative Binomial Approach. Transportation Research Record: Journal of the Transportation Research Board, No. 2236(05), p. 41–48.

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

Wanga, X., Chena, S. and Zheng, W., 2013. Traffic Incident Duration Prediction Based on Partial Least Squares Regression. Social and Behavioral Sciences, Volume 96, pp. 425-432.

Wang, et al., 2016. Identification of freeway secondary accidents with traffic shock wave detected by loop detectors. Safety Science, Volume 87, pp. 195-201.

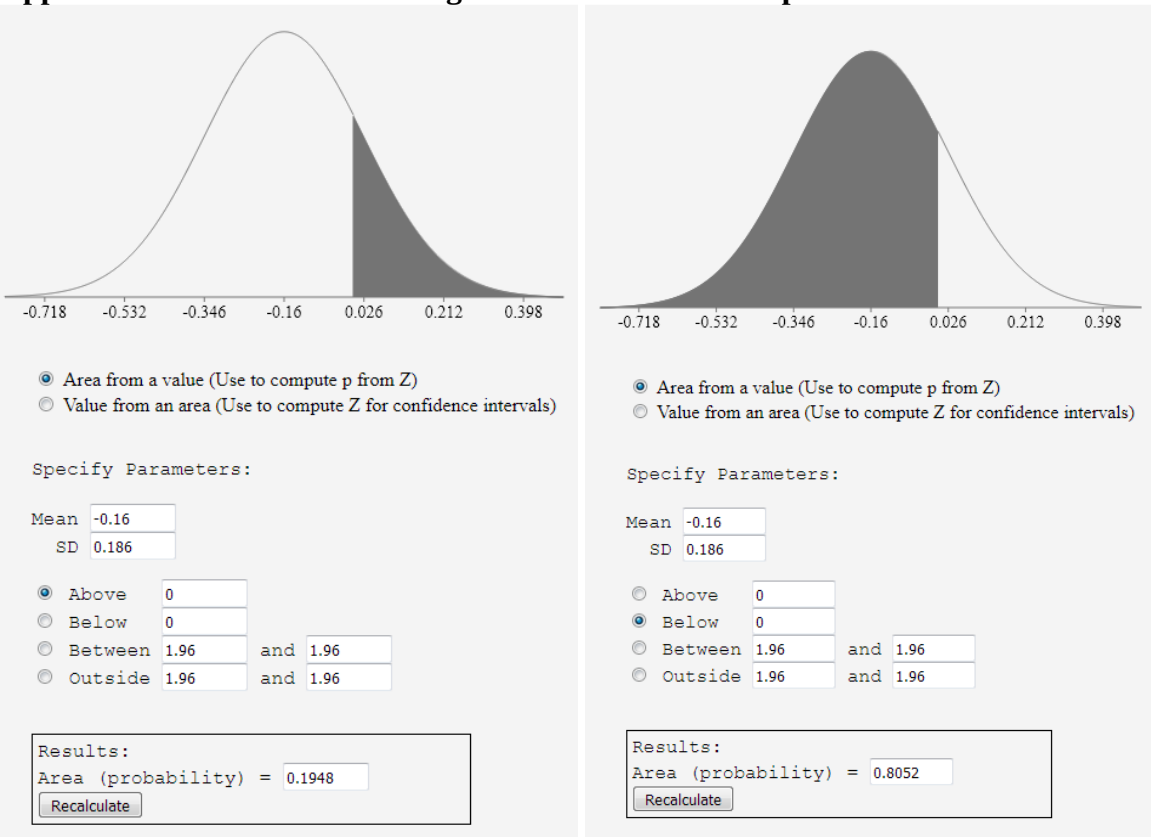
World Health Organization, 2015. Global Status Report on Road Safety, Geneva, Switzerland.: World Health Organization.

Xie, Y., Zhao, K. and N, 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis and Prevention, Volume 47, pp. 36-44.

4.9 Appendix 4A
Table 4A.1: Descriptions of the Significant Variables

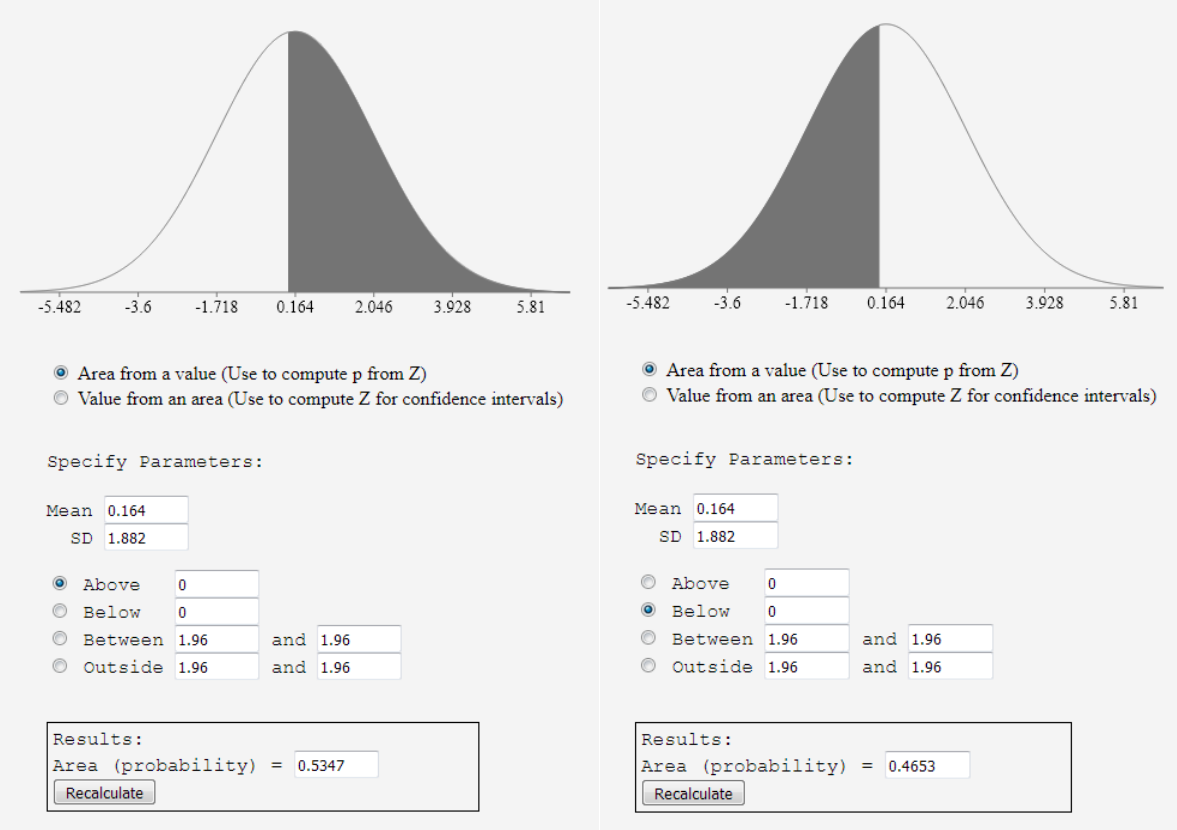
Category	Description
No congestion	
AADT	Traffic volume in 1000 vehicles
Dry road surface	1 if surface road condition is dry 0 otherwise
Urban area	1 if crash location is urban area 0 otherwise
Weekend	1 if day is Saturday or Sunday, 0 otherwise
Heavy trucks	Percent heavy vehicles
Low congestion	
Early morning	1 if time is between 1am and 5am, 0 otherwise
Misty weather	1 if weather condition is misty 0 otherwise
Medium congestion	
Rain	precipitation
AADT	Traffic volume in 1000 vehicles
Warning signs	1 if crash occurs outside of the warning signs area 0 otherwise
Vehicle towed	1 if vehicle involved required towing 0 otherwise
DUI	1 if the driver was under the influence of alcohol, 0 otherwise
Early morning	1 if time is between 1am and 5am, 0 otherwise
Winter season	1 if month is either December, January or February 0 otherwise
Bridge	1 if crash occurred on Feature on Bridge/Overpass/Underpass 0 otherwise
High congestion	
AADT	Traffic volume in 1000 vehicles
Speed	1 if speed is greater than 60mph, 0 otherwise
Work zone workers	1 if workers are present in the work zone 0 otherwise
Snow	1 if surface road condition is snow 0 otherwise
Morning peak time	1 if hour is from 0600hrs to 0800hrs 0 otherwise
Afternoon peak time	1 if hour is either 1600hrs or 1700hrs or 1800hrs 0 otherwise
Physical barrier present in median	1 if there is a physical barrier between opposing lanes 0 otherwise
Fatal Crash	1 if the crash severity is fatal, 0 otherwise

Appendix 4B- Estimation of marginal effects for random parameters



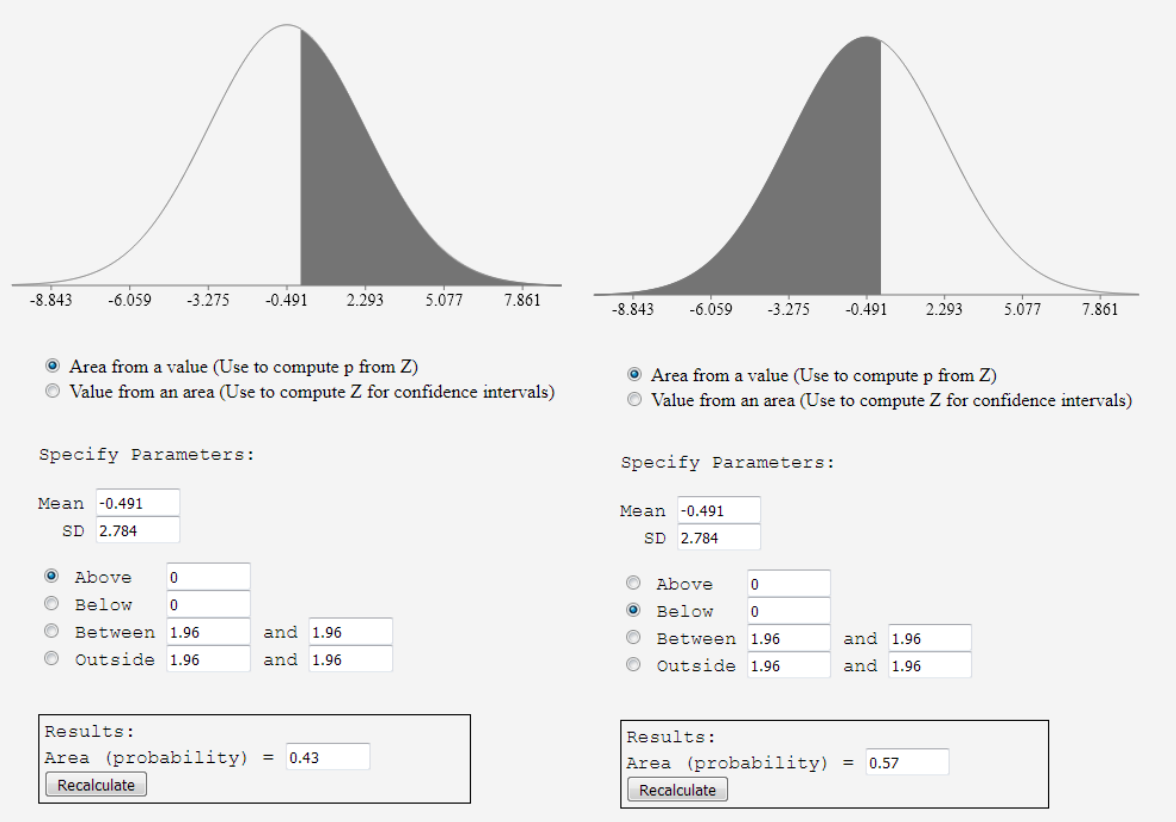
(a) (b)

Figure 4B.1: Estimating Random effects of heavy trucks
Source: http://onlinestatbook.com/2/calculators/normal.html



(a) (b)

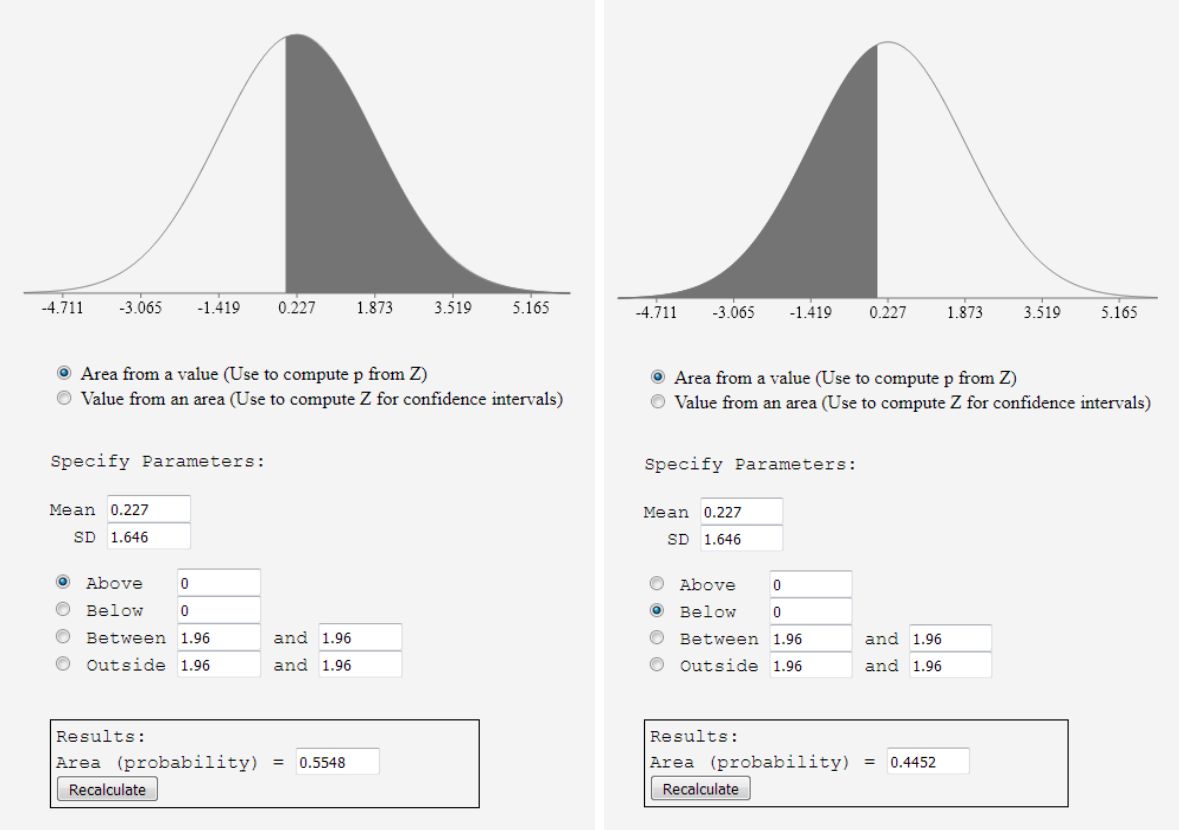
Figure 4B.2: Estimating Random effects of winter season
Source: http://onlinestatbook.com/2/calculators/normal.html



(a) (b)

Figure 4B.3: Estimating Random effects of bridge sites

Source: http://onlinestatbook.com/2/calculators/normal.html



(a) (b)

Figure 4B.4: Estimating Random effects of high severity crash
Source: http://onlinestatbook.com/2/calculators/normal.html

CHAPTER 5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This research demonstrates the importance of accounting for spatial features in crash analyses. Chapters 2 and 3 illustrated how spatial variations of socioeconomic factors affects traffic safety among and within different regions (i.e., different spatial scales). Chapter 4 shows that accounting for the spatial relationships (in conjunction with the time domain) between crash locations and the extent to which upstream traffic is impacted, reveal useful traffic management and incident response related information. Specific observations gleaned from the individual chapters are summarized in the following sections along with recommendations for future research.

The two studies on DUI (i.e., chapters 2 and 3) demonstrate the importance of accounting for spatial variation among socioeconomic factors. Both accentuate the need for further investigation of unobserved heterogeneities.

The Geographically Weighted Poisson Regression (GWPR) model identified the relationship between the DUI crashes and socioeconomic factors per region while taking care of spatial dependence among various postal codes. The findings indicate that there are four significant local variables with significant geographical variability which characterize DUI crashes in any postal code.

To begin with, *Rate of employment* is positively related to DUI crashes in most of the postal codes. Because of the geographic variability, there are some postal codes where rate of employment is negatively related to DUI crash frequencies. Employment significantly vary

across geography and it influence how people behave towards risk attributed to driving while drunk. Second, *percentage of people living in rented housing* has geographically varying impact on driver behavior. In general, findings show that DUI crashes are more in rural regions (where few people live in rental housing) than in urban regions (where more people live in rental housing). Third, *income* is negatively associated with DUI crash frequencies in most of the postal codes. The relationship is however different across different regions. These significant geographic variability is probably a function of the unobserved heterogeneities. Finally, *population density* was also found to have a positive association with DUI crashes in most postal codes. DUI crashes are generally high in postal codes with higher population density that in postal codes with low population density.

In addition, a suite of spatial econometric models was estimated for DUI crash rates per region in Alabama using macro level socioeconomic factors. Like the GWPR results, four categories of socioeconomic factors were found to influence the DUI crash rates. First, *Employment*- Three aspects of employment that influence driver behavior, particularly, DUI include the total rate of employment, employment rate for young people aged between 20 to 24 and employment rate for men. The research indicates that as total rate of employment increase, DUI crashes also increase. Similarly, an increase in employment rate of young people between ages 20-24, increases crash rates. Employment of more men however, reduced DUI crashes. These findings do not lead to recommendation that employment should be reduced, however, it points to the type of postal codes and people who should be targeted for education and increased awareness. Second, *Family and Housing* – This include average household size and percentage of people living in rented housing. Results indicated that an increase in household size reduce DUI crash frequency. Similarly, locations where most people lived in rental housing also have

fewer crash rates than locations where most people live in own housing. Which basically, show that there are more DUI crashes in rural regions than in urban regions. Thirdly, *Education* - higher education is negatively associated with DUI crash rates. Particularly, as more women get educated, DUI crash rates decrease. Finally, *Income* – on average as income increases, the DUI crash rates reduce in a postal code. This also concurs with the indication that rural regions (with low income) have high DUI crashes than urban regions (with high income).

Lastly, this research identified the relationship between crashes and highway traffic congestion by focusing on four categories of congestion outcomes following a crash event namely - No congestion, Low congestion, Medium congestion and High congestion. *No congestion* outcome following a crash event depends on many factors. The probability of having no congestion increases during weekend (Saturday and Sunday) and when the road surface is dry. However, the probability of having no congestion reduces in urban areas, and when there is an increase in AADT and particularly heavy truck volume. *Low congestion* on the other hand is expected to be low if a crash occurred between 1:00 am and 5:00 am or if weather was misty at the time of the crash. *Medium congestion* outcome followed a crash event during any of the following circumstance - rainy weather, outside of a warning sign area, if the vehicle in the crash needed to be towed, if the drivers in the crash was under influence of alcohol, if the crash occurred between 1:00am and 5:00 am or in winter (December, January, February), if the crash occurred at a section of a bridge and generally, if there is an increase in AADT. Finally, *High congestion* outcome is more likely if a crash event took place under any of the following general conditions - if AADT increases, if the speed at time of crash was higher than 60mph, if it occurred at a work zone, if there was snow, if it occurred during peak time (either morning or

afternoon peak time), for fatal crashes and if there is a physical barrier/median separating opposing traffic lanes.

5.2 Recommendations

Whereas the objectives of this research were successfully achieved, some recommendations are warranted as follows.

- Investigate socioeconomic influence on other driver-related crashes such as speeding, aggressive driving, distracted driving, seatbelt use, mechanical failure, etc.
- Use spatial results to drill down to deeper understanding of relationship between crashes and the types of people who cause them
- Investigate spatial relationships between where crashes occur and where at-fault drivers live
- Conduct spatial analyses at different scales (e.g., TAZ) to align with planning methods
- Investigate ways to leverage congestion severity work to understand occurrences of secondary crashes
- Examine effects of treating congestion severity (i.e., SDMH) variable as a continuous variable

REFERENCES

Adler, M, Ommeren, J. and Rietveld, P., 2013. Road congestion and incident duration, The Netherlands: VU University Amsterdam.

Anderson, T., 2007. Comparison of spatial methods for measuring road accident 'hotspots': a case study of London. Journal of Maps, 3(1), pp. 55-63.

Anderson, T, 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis and Prevention, Volume 41, pp. 359-364.

Bíl, M., Andrasi, R. and Janoska, k., 2013. Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accident Analysis and Prevention, Volume 55, pp. 265-273.

Black, W, 1991. Highway Accidents: A Spatial and Temporal Analysis. Transport Research Record, Volume 1318.

Borruso, G., 2008. Network Density Estimation: A GIS Approach for Analyzing Point Patterns in a Network Space. Transactions in GIS, 12(3), pp. 377-402.

Dickerson, A., Peirson, J. and Vickerman, R., 2000. Road Accidents and Traffic Flows: An Econometric Investigation. Economica, Volume 67, p. 101–21.

Drakopoulos, A., Shrestha, M. and Örnek, E., 2001. Freeway Crash Timeline Characteristics and Uses. Transportation Research Record: Journal of Transport Research Board, 1748(01), pp. 132-143.

Erdogan, S., 2009. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. Journal of Safety Research, Volume 40, pp. 341-351.

Erdogan, S., Yilmaz, I., Baybura, T. and Gullu, M., 2008. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accident Analysis and Prevention, Volume 40, pp. 174-181.

Flahaut, B., Mouchart, M., Martin, E. and Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones A comparative approach. Accident Analysis and Prevention, Issue 53, pp. 991-1004.

Fotheringham, A, Brunsdon, C. and Charlton, M., 2000. Quantitative Geography: Perspectives on Spatial Data Analysis. 1st Edition ed. SAGE Publications Ltd.

Garib, A., Radwan, A. and Al-Deek, H., 1997. Estimating Magnitude and Duration of Incident Delays. Journal of Transportation Engineering, 123(6), pp. 459-466.

Gundogdu, I., 2010. Applying linear analysis methods to GIS-supported procedures for preventing traffic accidents: Case study of Konya. Safety Science, Volume 48, pp. 763-769.

Hojati, A, Ferreira, L., Washington, S. and Charles, P., 2013. Hazard based models for freeway traffic incident duration. Accident Analysis and Prevention, December, Volume 52, pp. 171– 181.

Hojati, A, et al., 2014. Modelling total duration of traffic incidents including incident detection and recovery time. Accident Analysis and Prevention, June, Volume 71, p. 296–305.

Islam, S., Jones, S. and Dye, D., 2014. Comprehensive analysis of single- and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. Accident Analysis and Prevention, 67(2), p. 148–158.

Jones, B., Janssen, L. and Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention, 23(4), pp. 239-255.

Khan, G., Xiao Qin, P. and David A. Noyce, P, 2008. Spatial Analysis of Weather Crash Patterns. Journal of Transportation Engineering, 134(5), pp. 191-202.

Krishna Kumar, V., Pulugurtha, S. and Nambisan, S, 2005. Identification and Ranking of High Pedestrian Crash Zones Using GIS. International Conference on Computing in Civil Engineering 2005, ASCE.

Levine, N., Kim, K. and Nitz, L, 1995. Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. Accident Analysis and Prevention, 27(5), pp. 663-674.

Lord, D. and Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, Volume 44, pp. 291-305.

Mannering, F. and Bhat, C, 2014. Analytical Methods in Accident Research: Methodological Frontier and Future Directions. Analytic Methods in Accident Research, Volume 1, pp. 1-22.

Mannering, F, Shankar, V. and Bhat, C, 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research, Volume 11, pp. 1-16.

Mehta, et al., 2014. Analyzing Crash Frequency and Severity Data using Novel Techniques. A dissertation. [Online] Available at: http://acumen.lib.ua.edu/content/u0015/0000001/0001738/u0015_0000001_0001738.pdf [Accessed 26 10 2017].

Mehta, G. and Lou, Y., 2013. Modeling school bus seat belt usage: Nested and mixed logit approaches. Accident Analysis and Prevention, 51(1), p. 56– 67.

Mitchell, A., 2005. The ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics. ESRI Press.

National Highway Traffic Safety Administration, 2010. Traffic Safety facts. Research Note. [Online] Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811363 [Accessed 10 2017].

Ord, J. and Getis, A., 1995. Local Spatial Autocorrelation Statistics: Distribution Issues and an Application. Journal of Geographical Analysis, 27(4).

Pulugurtha, S, Krishna Kumar, V. and Nambisan, S, 2007. New methods to identify and rank high pedestrian crash zones: An illustration. Accident Analysis and Prevention, Volume 39, pp. 800-811.

Quddus, M, Wang, C. and Ison, S, 2010. Road Traffic Congestion and Crash Severity: Econometric Analysis Using Ordered Response Models. Journal of Transportation Engineering © ASCE, May 136(5), p. 424–435.

Steil, D. and Parrish, A., 2009. HIT: A GIS-Based Hotspot Identification Taxonomy. International Journal of Computer Applications (IJCA), 16(2).

World Health Organization, 2015. Global Status Report on Road Safety, Geneva, Switzerland.: World Health Organization.

Xie, Z. and Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Computers, Environment and Urban Systems, Volume 32, pp. 396-406.