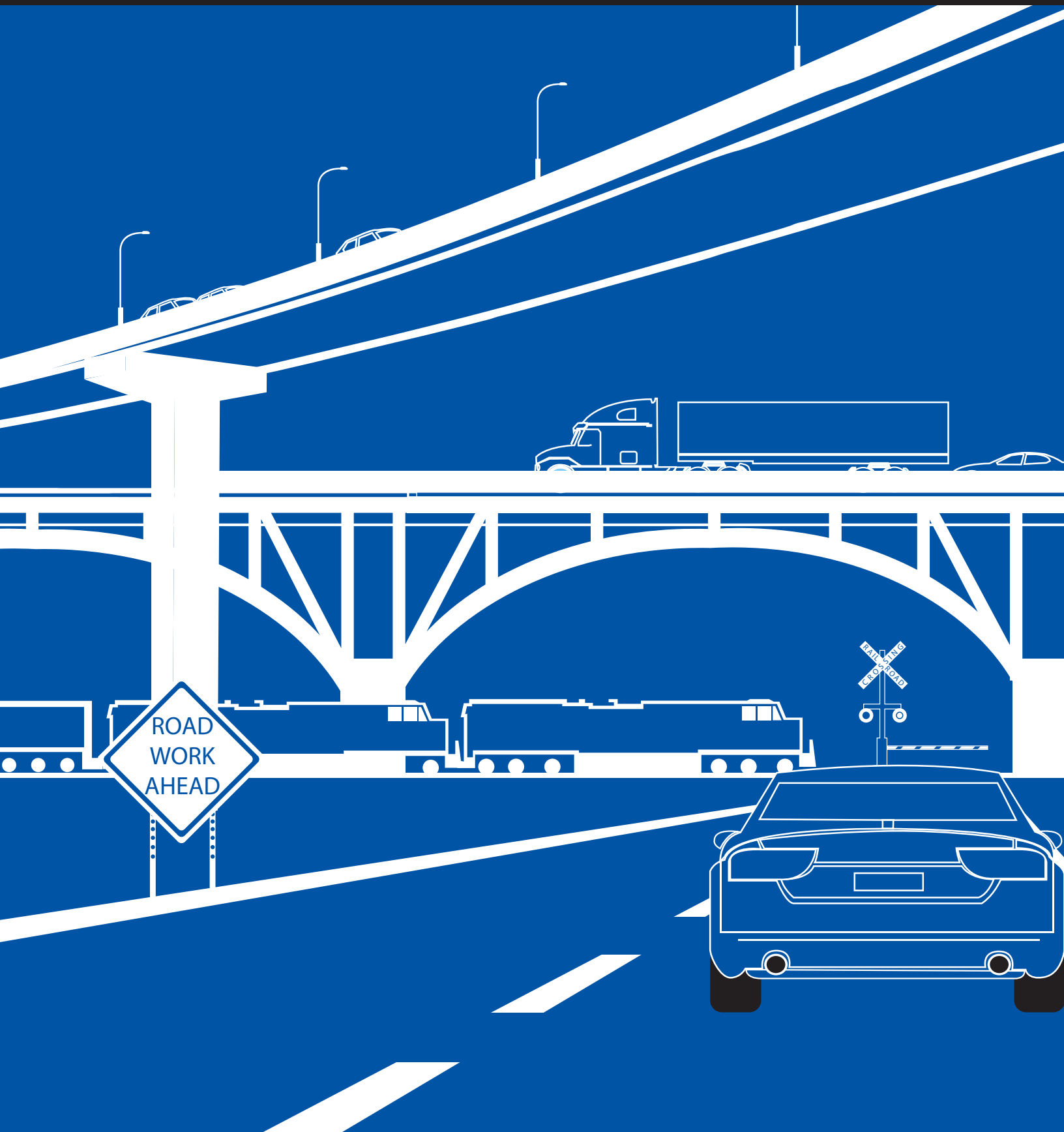
Excellence In Motion

Effect of Socioeconomic Factors on Kentucky Truck Driver Crashes

Report Number: KTC-20-21/KIPC20-1-1F

DOI: https://doi.org/10.13023/ktc.rr.2020.21



Kentucky Transportation Center
College of Engineering, University of Kentucky, Lexington, Kentucky

in cooperation with
Kentucky Transportation Cabinet
Commonwealth of Kentucky

The Kentucky Transportation Center is committed to a policy of providing equal opportunities for al persons in recruitment, appointment, promotion, payment, training, and other employment and education practices without regard for economic, or social status and will not discriminate on the basis of race, color, ethnic origin, national origin, creed, religion, political belief, sex, sexual orientation, marital status or age.

Kentucky Transportation Center
College of Engineering, University of Kentucky, Lexington, Kentucky

in cooperation with
Kentucky Transportation Cabinet
Commonwealth of Kentucky

KENTUCKY Transporation Center

Kentucky Transportation Center • University of Kentucky
176 Raymond Building • Lexington, KY 40506 • 859.257.6898 • www.ktc.uky.edu

University of Kentucky

Research Report
KTC-20-21/KIPC20-1-1F

Effect of Socioeconomic Factors on Kentucky Truck Driver Crashes

Nikiforos Stamatiadis, Ph.D.
Professor of Civil Engineering

Shraddha Sagar
Research Associate

Samantha Wright
Senior Lecturer in Civil Engineering

Aaron Cambron
Research Associate

Kentucky Transportation Center
College of Engineering
University of Kentucky
Lexington, Kentucky

June 2020

1. Report No. KTC-20-21/KIPC20-1-1F	**2. Government Accession No.**	**3. Recipient's Catalog No**
4. Title and Subtitle Effect of Socioeconomic Factors on Kentucky Truck Driver Crashes		**5. Report Date** June 2020 **6. Performing Organization Code**
7. Author(s): Nikiforos Stamatiadis, Shraddha Sagar, Samantha Wright, and Aaron Cambron		**8. Performing Organization Report No.** KTC-20-21/KIPC20-1-1F
9. Performing Organization Name and Address Kentucky Transportation Center College of Engineering University of Kentucky Lexington, KY 40506-0281		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** U60OH008483-15-00
12. Sponsoring Agency Name and Address Kentucky Injury Prevention Center 333 Waller Ave., Suite 242 Lexington, KY 40504		**13. Type of Report and Period Covered** Final **14. Sponsoring Agency Code**
15. Supplementary Notes Prepared in cooperation with the Kentucky Injury Prevention Center		

16. Abstract

Kentucky crash data for the 2015-2016 period reveal that per capita crash rates and increases in crash-related fatalities in the state outpaced the national average. To explain why the U.S. Southeast sees higher crash rates than other regions of the country, previous research has argued the region's unique socioeconomic conditions provide a compelling explanation. Taking this observation as a starting point, this study uses an extensive crash dataset from Kentucky to examine the relationship between highway safety and socioeconomic and demographic characteristics. Its focus is single- and two-unit crashes that involve commercial motor vehicles (CMVs) and automobiles. Using binary logistic regression and the quasi-induced exposure technique to analyze data on the socioeconomic and demographic attributes of the zip codes in which drivers reside, factors are identified which can serve as indicators of crash occurrence. Variables such as income, education level, poverty level, employment, age, gender, and rurality of the driver's zip code influence the likelihood of a driver being at fault in a crash. Socioeconomic factors exert a similar influence on CMV and automobile crashes, irrespective of the number of vehicles involved. Research findings can be used to identify groups of drivers most likely to be involved in crashes and develop targeted and efficient safety programs.

17. Key Words highway safety: socioeconomic factors; quasi-induced exposure; commercial motor vehicles		**18. Distribution Statement** Unlimited with approval of the Kentucky Injury Prevention Center	
19. Security Classification (report) Unclassified	**20. Security Classification (this page)** Unclassified	**21. No. of Pages** 53	**19. Security Classification (report)**

Table of Contents

1. Introduction 1
2. Literature Review 3
   2.1 Commercial Vehicle Driver Issues 3
   2.2 Socioeconomic and Demographic Variables 5
   2.3 Analysis Methods 8
   2.4 Summary 10
3. Research Methodology 13
   3.1 Socioeconomic Descriptor Factors 13
   3.2 Variable Selection Methods 14
      3.2.1 Correlation Test 14
      3.2.2 Recursive Partitioning Analysis 14
      3.2.3 Identifying Interactions 14
   3.3 Crash Exposure – Quasi-Induced Exposure Technique 15
   3.4 Statistical Modeling 15
      3.4.1 Assumptions 16
      3.4.2 Relative Accident Involvement Ratio 17
      3.4.3 Evaluation Criteria 18
   3.5 Model Development Approach 19
4. Data Collection and Preparation 20
   4.1 Crash Data 20
      4.1.1 Commercial Vehicle Crash Data 21
      4.1.2 Automobile Crash Data 23
   4.2 Socioeconomic Data 24
5. Statistical Modeling 27
   5.1 Commercial Vehicles 27
      5.1.1 Variable Selection 27
      5.1.2 Regression Models 29
   5.2 Automobiles 36
      5.2.1 Variable Selection 36
      5.2.2 Regression Models 38
6. Conclusions 43
References 45

List of Tables

Table 1 List of Crash Record Variables 20
Table 2 Commercial Vehicle Driver Age Distribution (2013-2016) 22
Table 3 Human Factor Code Frequency for Commercial Vehicle Crashes (2013-2016) 23
Table 4 Automobile Driver Age Distribution (2013-2016) 24
Table 5 Sample Size of the Datasets 24
Table 6 List of Socioeconomic Variables 25
Table 7 Correlation Analysis Results for Commercial Vehicle Crashes 28
Table 8 Models for Single-Unit Commercial Vehicle Crashes 30
Table 9 Statistics of Single-Unit Commercial Vehicle Models 31
Table 10 Ratios of Single-Unit Commercial Vehicle Model 31
Table 11 Ratios of Single-Unit Commercial Vehicle Model (Age Groups Only) 32
Table 12 Human Factors Model for Single-Unit Commercial Vehicles 32
Table 13 Statistics of Human Factors Model for Single-Unit Commercial Vehicle 33
Table 14 Model for Two-Unit Commercial Vehicle Crashes 33
Table 15 Statistics of Two-Unit Commercial Vehicle Models 34
Table 16 Ratios for Two-Unit Commercial Vehicle Model 35
Table 17 Human Factors Model for Two-Unit Commercial Vehicles 35
Table 19 Correlation Analysis Results for Automobile Vehicle Crashes 37
Table 20 Model for Single-Unit Automobile Crashes 39
Table 21 Statistics of Single-Unit Automobile Models 40
Table 22 Ratios of Single-Unit Automobile Model 40
Table 23 Models for Two-Unit Automobile Crashes 41
Table 24 Statistics for Two-Unit Automobile Vehicle Models 42
Table 25 Ratios for Two-Unit Automobile Vehicle Model 42

List Of Figures

Figure 1 RAIR for Driver Gender 18

1. Introduction

According to World Health Organization (WHO), every year 1.25 million people die in road traffic crashes, while at least 20 million are involved in non-fatal crashes (WHO 2018). In the United States (U.S.), road traffic crashes are a leading cause of death. Crash data from the Kentucky Transportation Center (KTC) indicate that fatalities increased from 761 to 834 between 2015 and 2016. An increase of 10 percent, this was higher than the national average (KTC 2016). In addition, Kentucky has a higher overall crash rate per population than the national average. In 2016, the National Highway Traffic Safety Administration (NHTSA) estimated 22.5 crashes per 1,000 persons at the national level, while Kentucky had a rate of 37.3. NHTSA also estimated the value of societal harm — which includes economic impacts and a valuation for lost quality of life — for all traffic crashes in 2010 at $836 billion (NHTSA 2015).

Of the 269 million registered vehicles in the U.S. in 2016, approximately 11.5 million were commercial vehicles (trucks and buses) according to the Federal Motor Carriers Safety Administration (FMCSA 2018). Approximately 6.1 million drivers hold a Commercial Driver License (CDL), comprising only 3 percent of the overall number of licensed drivers. Commercial vehicles (CMV) were involved in 4,079 fatal crashes in the U.S. in 2016, representing 11.8 percent of the total number of fatal crashes and 7.4 percent of non-fatal crashes. Even though CMVs typically drive longer distances, their crash involvement could be considered somewhat proportional to their exposure. In 2016, the vehicle-miles of travel (VMT) for CMVs represented 9.1 percent of total VMT in the US (FMCSA 2018). However, the number of CMV crashes has increased constantly since 2009. In 2017, the number of people who died in large truck crashes was 30 percent higher than in 2009, according to the Insurance Institute for Highway Safety (IIHS; 017). Most deaths in crashes involving CMVs are passenger vehicle occupants, which is mainly attributed to the vulnerability of people traveling in smaller vehicles.

Several factors could be associated with roadway crashes and their occurrence. Addressing issues that could lead to safety problems, would improve overall roadway safety. It is therefore important to understand the underlying factors contributing to crashes to implement effective countermeasures to improve traffic safety. In several such attempts, driver behavior, demographic factors, socioeconomic features, geometric design and roadway characteristics were identified as associated factors (Noland et al. 2004; Hasselberg et al. 2005; Aguero-Valverde et al. 2006; Factor et al. 2008; Hanna et al. 2012; Brown 2016; Adanu et al. 2017; Zephaniah et al. 2018). Past research efforts demonstrated significant influence of macro-level socioeconomic features on crash occurrence (e.g., poverty, income, employment and education) (Stamatiadis et al. 1999; Factor et al. 2008; Brown 2016; Kocatepe et al. 2017; Zephaniah et al. 2018). Many of these studies concentrated on the socioeconomic factors of the region where the crash occurred. Maciag (2014) compiled fatal pedestrian accidents reported in Fatality Analysis Reporting System (FARS) for the 2008 to 2012 period to study the relationship between fatal crashes and the economic condition of the crash location. He found that fatalities are generally more common in poor areas. Also, an historical crash data analysis by NHTSA indicated that crash rates are 2.5 times higher in rural areas than in urban areas (NHTSA 2013). These studies underscore the greater potential for crashes in socially and economically disadvantaged areas. Though it is important to examine the socioeconomic characteristics of the region where a crash occurred, it may provide more information to focus on the residence characteristics associated with the drivers who cause a crash.

Prior research attempted to determine the association between socioeconomic factors related to driver residence and crash occurrence and estimate their role in crash occurrence (Blatt and Furman 1998; Chandraratna 2004; Chandraratna et al. 2005; Lee et al. 2014). A recent WHO study found that people of lower socioeconomic backgrounds are more likely to be involved in road traffic crashes, with causation factors including human errors, such as speeding, lack of restraints, distracted driving, driving under the influence, inadequate roadway infrastructure, and traffic law enforcement (WHO 2018). Blatt and Furman

also reached the same conclusion, correlating socioeconomic characteristics of driver residence with crash occurrence (Blatt and Furman 1998). They demonstrated that fatal crashes are more likely to take place on rural roads, while drivers who reside in rural areas or small towns have significant involvement in these crashes. Several other studies have confirmed the high risk of crash involvement for drivers residing in a rural/poor neighborhood (Lee et al. 2014; Brown 2016; Ivan et al. 2016; Zephaniah et al. 2018).

Stamatiadis and Puccini (1999) showed that the U.S. Southeast experiences consistently higher fatality rates compared to other regions. They noted that the distinct socioeconomic characteristics of the region could help explain the high fatality rate. They identified potential socioeconomic factors that could explain the high fatality rates in these regions, including median household income, unemployment levels, educational attainment, and percentage of rural population. The study suggests the socioeconomic data for a driver's residence zip code could serve as a potential surrogate measure for explaining high fatality rates.

A plausible explanation for increased crash rates in Kentucky may be the differences in a variety of socioeconomic characteristics of the state compared with other states. Based on statistics from the Bureau of the Census, Kentucky has lower percentages of high school completion and university attainment than the national average (U.S. Census Bureau 2018). With respect to income characteristics, most counties have a median family income 19 percent lower than the national median income, are at the bottom of the national rankings with respect to both income and disposable income per capita and have one of the highest percentages among all states of people living in poverty. These types of socioeconomic characteristics could influence highway safety by affecting the age of vehicles owned (older, less safe vehicles), vehicle condition (not properly maintained), the attitudes of the drivers toward safety and risk-taking behaviors, and the level of driving education available to people (Stamatiadis and Puccini 1999). Moreover, Kentucky is considered a rural state since more than fifty percent of its counties are classified as rural (Harrah 2015).

The number of traffic collisions is gradually increasing in Kentucky. Being a state in the Southeast, Kentucky's socioeconomic factors are suspected to be a significant reason for these recent increasing crash trends. In 2016, 834 fatalities were reported in Kentucky, a 9.6 percent increase over the previous year (KSP 2016). An increasing trend has been observed in commercial truck crashes since 2009. In the 2009-2016 period, CMV crashes increased 27 percent (KSP 2016). It is apparent that there may be some connection between socioeconomic factors and crash occurrence. Therefore, it is critical to examine their impact on each other. It is also important to determine how demographic data attributed to the socioeconomic characteristics of the driver residence influence crash involvement. Analyzing these factors might help identify the major drivers of increasing crash trends, and in turn, identify areas that may require additional attention for improving overall roadway safety.

Therefore, the primary goal of this research is to define at-risk driver groups based on the socioeconomic characteristics of driver residence, with an emphasis on CMV drivers. This study mainly analyzes the socioeconomic and geodemographic factors of driver residence zip code utilizing historical crash data from Kentucky. The main objective of the study is to identify factors that could potentially be indicators of CMV crash occurrence. The findings of this work will help identify groups of drivers with a high crash involvement risk factor. It is important to determine whether these drivers, who may contribute to the future crash risk, belong to a particular group (e.g., age, gender) or region (e.g., rural/urban). This would provide better evidence for implementing efficient safety programs that target such groups.

2. Literature Review

Much research effort has been undertaken globally to investigate the role and possible contribution of socioeconomic and demographic factors to crash occurrence. Some researchers have investigated demographics of areas surrounding the crash location, while others have used surrogate descriptors associated with residence location of drivers involved in a crash. The following sections discuss past research efforts which have focused on identifying significant socioeconomic and demographic factors that help explain crash involvement as well as methods used to investigate them. Moreover, CMV-related safety issues are presented to better understand what role they play in crashes and determine whether any occupational characteristics can be identified that would support the crash analysis being undertaken.

2.1 Commercial Vehicle Driver Issues

Considerable research has been done around the world to investigate the effects of vehicle, roadway, and driver characteristics on crash severity and occurrence. Notably, the role of alcohol and drug use has become a topic of interest, especially in relation to CMV drivers. Some analysis has focused on the prevalence of drug use among CMV drivers and related contributing factors, while other research has sought to determine the correlation between substance use and crash occurrence and severity. Prior research that examined factors which contribute to CMV crashes, and specifically the role that alcohol and drug use plays in those crashes, is presented here.

Several research efforts have focused on the factors that contribute to CMV drivers using some form of impairing substances, regardless of their effects on crash occurrence or severity. Girotto et al. (2014) and Knauth et al. (2012) examined the factors that contribute to CMV driver drug usage. Girotto et al. (2014) looked specifically into factors influencing the intake of psychoactive substances by synthesizing cross-sectional studies across multiple countries, with most of the data sourced from self-reporting surveys. They found that even though methodologies varied between studies, the intake of psychoactive substances is a relatively frequent occurrence — 19.3 percent of drivers admitted to having used marijuana during their lifetimes. They also concluded that poor working conditions, such as resting time, trip length, and night shifts, tended to increase the frequency of drug intake. Knauth et al. (2012) also relied heavily on survey data to analyze the factors affecting substance use among truck drivers in Brazil. The study focused on substances used with the intent of staying awake and used a Poisson regression to analyze the data. They found that approximately 23 percent of drivers used some substance to stay awake, with the majority using some form of amphetamines, while the consumption of alcohol on a weekly basis was reported by 45 percent of drivers. They also noted that longer trips along with younger age, higher income, and alcohol consumption were all associated with more frequent use of substances to stay awake. Like Girotto et al. (2014), they theorized that working conditions and the inherent physical and emotional stress of truck driving are major factors which lead to more frequent substance use.

Souza et al. (2005) focused primarily on the prevalence of sleep disorders and sleeping habits among truck drivers, but were able to discover through questionnaires that substance use is frequent among Brazilian truck drivers, with 95 percent reporting alcohol use and 11 percent reporting amphetamine use, of which 77 percent used more than six times a week. Similarly, Mir et al. (2013) interviewed 857 truck drivers in Pakistan about socioeconomics, sleep habits, drug use, and crash involvement. They discovered that 23 percent admitted to smoking marijuana while driving, 6 percent reported drinking alcohol right before driving, and 8 percent reported using stimulant pills. However, using multivariate logistic regression, only alcohol was found to have a significant effect on crash occurrence. It is important to note that all of these studies have been primarily based on self-reported data, and that introduces implicit bias into data on substance use frequency and the crash frequency.

To counter this, many researchers have conducted crash analyses using historical crash records and instances of drug and alcohol use recorded therein. The National Transportation Safety Board (NTSB; 1990) attempted to examine and identify the contributing factors to crashes looking only at crashes which were fatal to truck drivers. This approach resulted in a much higher reporting rate for toxicological samples indicating alcohol and drug use. The NTSB studied fatal crashes across eight states for 1988 using FARS crash data and contracted with these states to obtain consistent toxicological tests for these fatal crashes. Alcohol and marijuana were both involved in 13 percent of fatal truck driver crashes. In addition, caffeine was involved in 35 percent, with other drugs and amphetamines less common. While this information indicates the presence of substance use, it only does so for drivers involved in fatal crashes, reducing the sample size significantly and possibly underestimating the level of use.

Chen et al. (2015) analyzed crashes on a slightly larger scale, considering two years of crash data for injury or fatal CMV crashes. This study used a Bayesian random intercept model and discovered that along with factors such as road grade, number of vehicles, and seatbelt usage, the presence of alcohol or drugs positively correlated with crash severity. However, crash severity does not necessarily indicate whether the truck drivers are driving less safely due to the presence of drugs. Gates et al. (2013) attempted to quantify this in relation to stimulant (e.g., amphetamines, cocaine) use by analyzing unsafe driving actions (UDAs) recorded in the FARS database. Drivers in this analysis were only included when their blood alcohol level tests were available. A logistic regression was used to calculate odds ratios and results indicated that stimulant-positive drivers had 78 percent greater odds of committing a UDA compared to drivers who were stimulant-negative. Khorashadi et al. (2005) also found that alcohol or drug use was the prominent causal factor in 4 percent of crash events for rural road truck crashes and the probability for severe/fatal injury increased 246 percent compared to crashes not involving alcohol or drugs. Conversely, Lemp et al. (2011) found that truck drivers under the influence of illegal drugs were most commonly involved in minor injury crashes as opposed to major injury or fatality crashes. While these studies elaborate on the relation between truck crashes and drug/alcohol use, they also rely on crash reports for information on drivers under the influence, which can result in problems such as sample size and collection bias. To adjust for a small sample size of vehicles actually tested for drugs (in this case marijuana), Chen et al. (2018) developed a multiple imputation procedure for estimating marijuana positivity among drivers with missing marijuana test results, using a Bayesian multilevel model that allows a nonlinear association with blood alcohol concentrations (BACs), accounts for correlations among drivers in the same states, and includes both individual-level and state-level covariates. The resulting adjusted marijuana positivity rate of 11.7 percent was lower than the observed rate of 14.8 percent among drivers involved in fatal motor vehicle crashes. Chen suggested that the multiple imputation model can reduce bias and improve efficiency in estimating positivity rates of marijuana.

The research presented thus far has primarily shown that truck drivers are at risk for alcohol and drug use, both during their lifetimes and on the job, and that substance use seems to correlate positively with both truck driver injury severity and crash occurrence. This does not however assess the risk to other drivers. Several researchers have investigated both the occurrence of truck driver fault in crashes and the differences in truck driver single crashes versus multi-vehicle crashes. Vachal (2016) investigated the difference in contributing crash factors as they differ among single and multivehicle truck crashes, finding that non-truck drivers who were involved in injury crashes with trucks had the highest rate of alcohol or drug usage at 16.6 percent, compared to 12.5 percent for drivers in non-truck related injury crashes, and only 1.3 percent for truck drivers involved in injury crashes. In addition, for multivehicle truck crashes involving alcohol or drugs, the risk of serious injury was twice as likely for the driver of the passenger vehicle compared to the truck driver. This research seems to indicate that while drugs and alcohol are potentially a contributing factor for truck drivers, substance use is more common and more dangerous for drivers of passenger vehicles. Chen et al. (2011) also separated crash analysis into single and multivehicle truck crashes. Using Highway Safety Information System (HSIS) data and multinomial logit models, they discovered that drivers influenced by drugs or alcohol were more likely to experience of injury or fatality according to the

multivehicle model, but this relation was not significant in the single vehicle model. This indicates that truck drivers under the influence of alcohol or drugs are the most dangerous when combined with the presence of other vehicles.

Finally, Spainhour et al. (2005) performed an in-depth analysis of fatal truck crashes by utilizing crash records, video logs, photographs and site visits. They found that drugs and alcohol use was much more common among truck drivers that were considered at-fault in fatal crashes. However, truck drivers were only at fault for approximately 30 percent of the crashes they had been involved in.

2.2 Socioeconomic and Demographic Variables

Various socioeconomic and demographic variables have been examined in the past to identify their potential contribution to crash rates. Prior research shows some common threads among explanatory variables, confirming a priori expectations: income, poverty, employment, education, rurality, and driver age all seem to have an impact (Stamatiadis et al. 1999; Noland et al. 2004; Factor et al. 2008; Lee et al. 2014; Brown 2016; Zephaniah et al. 2018).

Rural areas are generally cited as having higher fatality crash rates than urban areas and a large portion of previous research dealt with the levels of rural and urban components of a region. Muelleman and Mueller (1996) investigated the relationship between fatal CMV crash characteristics and population density. Information on human (age, gender, restraint use, alcohol, ejection from vehicle, seating position, driving record), vehicle (vehicle make, crash type, manner of leaving scene, most harmful event), and crash variables (crash location, crash time, posted speed limit, first harmful event, surface type, emergency medical system (EMS) times) were included in the analysis. The counties in the study regions were categorized as urban and rural; rural counties were further subdivided into three groups based on population density. Major factors that were significantly related to high fatality rates in low density areas were prevalence of alcohol use and higher level of intoxication, delayed medical care, use of light and heavy trucks, frequent non-collisions (defined as a crash with no injuries or damages) on less-travelled roads, and frequent crashes on gravel-surfaced roads. Also, the study confirmed the previously identified inverse relationship between population density and MVC fatality rates. They concluded that the fatality rate per 100 million VMT was 44 percent higher in rural than urban areas and noted that rural areas are not homogeneous, which comparisons based only on urban/rural groupings can obscure. However, variables like restraint use, crash severity, and older occupants showed no difference between the three rural regions, raising concerns about their role in explaining the relationship between fatality rate and population density. Though this research recognized many crash variables associated with population density, it did not determine the relative contribution of each factor for explaining differences in fatality rates in rural areas. The authors recommend further research to determine how the fatality rate increase in areas with low population density is associated with pre-crash, crash, and post-crash variables. However, there has not been relevant research conducted in these areas yet.

Blatt and Furman (1998) conducted a similar geodemographic analysis at the zip code level, with a focus on the residential location of the driver, which was classified as rural and urban. Five levels of population density were identified for classifying each driver's residence location: rural, small town, second city, suburban, and urban. Other driver characteristics were divided into social clusters (age groups, gender, involvement in crash resulting death of a child, blood alcohol concentration level). Using geodemographic analysis, the percentage of drivers in fatal crashes in each social cluster was compared to the base population of that social cluster. Findings indicated that drivers from rural areas or small towns were more likely to be involved in fatal crashes and that those fatal crashes were more likely to take place on rural roads. They also acknowledged that roadway features (e.g., two-lane highways, narrow shoulder, limited sight distance) may play a bigger role in rural crashes while economic and behavioral factors (e.g., seat belt use, poor EMS response time, longer travel time to reach the nearest medical facility) could contribute to serious crash outcomes. Zwerling et al. (2005) investigated factors associated with increased fatal crash involvement

rates in rural areas. They found that fatal crash incidence density was more than two times higher in rural than in urban areas. The major reason was the high rate of increased injury severity in rural crashes — three times higher in rural areas compared to that in urban areas.

Noland and Quddus (2004) used negative binomial (NB) regression to explore the association between crash casualties and land use variables (proportion of urbanized area, population, employment density), road characteristics (length of road types, number of junctions and roundabouts) and areawide demographics features (age, level of social deprivation, percent of economically active population). NB models were developed for total fatalities, serious injuries, and minor injuries. The results showed that densely populated urban areas had fewer traffic causalities, while areas with higher employment saw more traffic causalities. The roadway characteristics examined as part of the study did not exhibit any influence on traffic casualties, although the length of the road segments showed some effect on serious injuries. Social deprivation showed a positive relationship with traffic causalities but no significance for motorized (excluding bicyclists and pedestrians) casualties. Also, the residual cause for high causality rate in areas with higher levels of social deprivation was not investigated. They offered as a possible explanation for their findings the notion that lower income people tend to live in areas with low cost of living and cheap housing, while such areas are likely to have unsafe roadway conditions. Reviewing this issue would be useful to identify target areas or populations that need more attention.

Hasselberg et al. (2005) determined that drivers with a relatively low educational attainment level showed excess risk for overall crashes and crashes leading to fatality or serious injury. Their study also estimated that 33 percent of minor injuries and 53 percent of severe injuries would be avoided if all subjects had the same injury rate as subjects with a higher education. Similarly, Zephaniah et al. (2018) revealed that DUI crash rates (normalized by population) were influenced by employment, income, education, and housing characteristics. Areas with high rental housing percentages exhibited lower DUI crash rates. The rate of DUI crashes was higher in rural areas, possibly indicating acceptance of drunk driving among communities living in those regions. Also, the overall percentage of residents with at least a high school education in a postal code reduced the number of DUI crashes. Their study also showed that DUI crashes were related to lower male employment and female educational achievement, while Cook et al. (2005) also confirmed higher DUI crash involvement for male drivers. These studies used the characteristics of the driver's residence location and showed that greater educational attainment reduces on vehicle crashes.

Both income and poverty have been cited as relevant predictors for use in crash-related analysis in several sources. However, income and poverty could be closely related as poverty status is generally based on income below a certain level. Lee et al. (2014) investigated the relationship between at-fault driver residence characteristics and all types of crashes for three years of data in Florida. They found that median family income had a negative relationship with the number of at-fault drivers, indicating that drivers from lower income communities are more likely to be responsible for a crash occurrence. Maciag (2014) indicated that within metro areas, low-income tracts recorded pedestrian fatality rates approximately twice that of more affluent neighborhoods; high poverty rate tracts revealed a similar trend. Aguero-Valverde et al. (2006) also concluded that the percentage of the population living in poverty had a highly significant and positive correlation with crash risk when using a NB prediction model.

Employment has been cited in several forms either as unemployment rates, portion of people working from home, or portion of unskilled workers. Factor et al. (2008) used a sample of the Israeli population with detailed socioeconomic data and nine years of crash data for their analysis. They found that non-skilled workers were over-involved in fatal crashes relative to their size in the total population of all workers. Conversely, Lee et al. (2014) found that the higher proportion of the population working from home resulted in a lower number of at-fault drivers, though it was proposed that this is the result of travel exposure. Later, Adanu et al. (2017) found that unemployed drivers had probability of 0.23 of being at-fault in a crash while the probability of being at-fault in a serious injury crash was 0.57. They suggested that the odds of an

unemployed driver being at-fault for a serious crash were 1.32 times higher than for a driver who was employed, self-employed, or retired. In addition to employment, they attempted to demonstrate that average credit scores (lower scores equal higher risk) and average commute times (longer times equal higher risk) are significant predictors for severe injury crash risk. At the driver level, a significant proportion of serious injury crashes involved no seat belt usage; other contributing factors included employment status, driver age, distracted driving, and the driver's race. The model also verified a previously known inverse relationship between population density and severity of crashes, however, the authors counterintuitively suggested that larger populations are more likely to live in urban areas having higher overall incomes and educational levels, which are factors that may influence crash occurrence and severity. Even though the influence of population density and vulnerability of rural areas on severe crashes have been established by previous studies, the authors suggest that more detailed investigation of less populated regions are needed to better understand the relationship between driver characteristics and specific crash types.

Age is a predominant demographic phenomenon that contributes to a driver's involvement in a crash. Brown et al. (2016) attempted to identify and analyze the socioeconomic and demographic factors related to the residential characteristics (at zip-code level) of drivers involved in crashes. Their study found drivers aged 15-19 had the highest odds of being at risk for an injury or fatal crash, followed by the 20-24 age group. Middle-aged drivers (45-54) had the lowest odds of being at-fault in a crash. Chen et al. (2010), Factor et al. (2008) and Hanna et al. (2012) all demonstrated that undesirable crash results, such as more crashes or higher fatality rates, were present for young or new drivers, but there was some variation about the impact of elderly drivers. It might be that the young drivers have a tendency to speed more than older drivers (NHTSA 2013). Lee et al. (2014) determined that a larger proportion of elderly population reduces the likelihood of drivers being at fault. Also, Adanu et al. (2017) indicated that older drivers (above 65 years) had the lowest contribution to fatal crashes. This might be because older drivers contribute less to a region's socioeconomic features (e.g., median income) than other age groups. Males (2009) showed a joint effect of age and income-related factors that contributed to young driver fatalities. Using a multivariate regression analysis, he concluded that driver age was not a significant predictor of fatal crash risk when poverty-related factors (e.g., older vehicle age, lower state per capita income, and lower education levels) were controlled. Aguero-Valverde and Jovanis (2006) found that counties with a higher percentage of the population living in poverty; a higher percentage of their population in age groups 0–14, 15–24, and over 64; and those with increased road mileage and road density experienced significantly increased crash risk. Several studies of older adult drivers discussed the risk factors they create for themselves and others. Lyman et al. (2002) observed increasing fatal crash rates for drivers over the age of 70. The study found that drivers over 65 years of age will account for more than half of the total increase in fatal crashes by 2030. However, the contribution of different age groups to crash severity is still not clear and should be investigated further.

In addition to age, crash occurrence is often associated with the gender and marital status (separated or widowed) of the driver (Factor et al. 2008; Kocatepe et al. 2017). Factor et al. (2008) provided evidence that separated and widowed drivers are 50 percent more likely to be involved in a crash than married drivers. In terms of the at-fault driver, the proportion of males was higher than that of females. For the state of Kentucky, 55 percent of the drivers who were involved in collisions during 2016 (where the gender was listed) were male while 45 percent were female. In fatal collisions, 74 percent of the drivers were male, and 26 percent were female. Zephaniah et al. (2018) showed that DUI crashes related to male employment and female education attainment. Additionally, there might be a joint relationship between other socioeconomic factors (like income) and gender and age which requires more investigation.

Another interesting factor contributing to crash occurrence is the proximity to driver residence (Brown 2016; Adanu et al. 2018). A latent class analysis (a model-based clustering method) conducted by Adanu et al. (2018) illustrated that over 75 percent of young at-fault driver crashes occurred within 25 miles of the driver's residence. However, Brown (2016) showed that approximately 35 percent of the crashes occurred within 5 miles of the driver's residence. Additional investigation is recommended into how crash

occurrence is influenced by the proximity to driver residence belonging to specific target group (e.g., age, gender, educational attainment, or regions, rural/urban area).

Apart from socioeconomic and demographic characteristics of the driver's residence, driver history plays a major role in future crash risk. Chandraratna (2005) demonstrated that a driver who had one previous at-fault crash was about 150 percent more likely to be involved in a another crash than a driver who had no previous at-fault crash involvements. His study also demonstrated that drivers who have driving records with citations, crashes, or both were high-risk drivers. Even though his research estimated the likelihood of a driver being involved in a future crash, the forecast was limited to at-fault drivers with previous crash records. Many researchers have investigated the impact of crash-prone drivers on safety and developed models predicting how a driver's crash history could affect their crash occurrence(s) in the upcoming year (Blasco et al. 2003; Sun et al. 2014).

2.3 Analysis Methods

The NB distribution is a discreet probability distribution often used when dealing with crash counts. NB regressions are used to model crash counts for a roadway segment. Noland and Quddus (2004) used NB count data models to analyze the associations between demographic factors (e.g., land use types, road characteristics, areawide demographics, including the level of social deprivation) with traffic fatalities and serious or slight injuries. Social deprivation is an index developed in the United Kingdom consisting of six socioeconomic factors: income, employment, health deprivation and disability, education skills and training, housing, and geographical access to services. They used the census block in England as a spatial unit for crash locations to relate these demographic characteristics to crash fatalities. More recently, the *Highway Safety Manual* (HSM) recommended developing Safety Performance Functions (SPFs) using NB regressions which are primarily based on Average Annual Daily Traffic (AADT) for homogeneous roadway segments. However, Ivan et al. (2016) demonstrated an alternative for predicting crashes on local roads if traffic volumes are not available. The study estimated SPFs for local road intersections and segments at the Traffic Analysis Zone (TAZ) level using sociodemographic and network topological data. There are approximately 1,800 TAZs in Connecticut, which were then clustered into six analysis groups based on land use and population density. SPFs were developed using Poisson regression models, which can predict intersection and segment crashes in each TAZ using the number of intersections and the total local roadway length, respectively.

Other forms of regression modeling have been used in crash analysis. La Torre et al. (2007) and Rivas-Ruiz et al. (2007) used multiple linear regression in their analysis while Chen et al. (2015) used a Bayesian random intercept regression model. La Torre et al. (2007) investigated the association between regional differences in traffic crash mortality and crash rates and sociodemographic factors and variables describing road behavior, vehicles, infrastructure and medical care in Italy. Rivas-Ruiz et al. (2007) used simple and multiple linear regression with a backwards stepwise elimination approach to study the variability of Road Traffic Injury (RTI) mortality on Spanish roads, adjusted for Vehicle Kilometers Traveled (VKT) in each Spanish province. Both studies found areawide socioeconomic factors (e.g., employment rates, alcohol use, education levels) to be somewhat significant. Chen et al. (2015) analyzed injury or fatality truck driver crashes. The study concluded that the presence of alcohol or drugs had a positive correlation with crash severity.

Some have found other regression models to be more useful, such as logistic and lognormal regressions. Logistic regression is the simplest type of regression that can be used when the dependent variable is binary. This technique fits the best when the effect of more than one independent variable (categorical, continuous, or both) is examined. Factor et al. (2008) created a binary response variable to describe crash fatality level. The model used demographic factors to predict the probability of being involved in a fatal crash versus a non-fatal crash. The research linked nine years of injury and fatal road-crash records with census data and used several socioeconomic factors all grouped into discrete categories, such as gender, education groups,

and age groups. The binary dependent variable indicated whether the driver had been involved in a fatal or severe accident within the past nine years. They also used categorical independent variables such as gender, age groups and marital status for analysis. The findings of the regression were turned to probabilities, which is one of the major contributions of logistic regression. Vachal (2016) used logistic regression to study crash factors in relation to injury outcomes for single and multivehicle truck crashes, finding that while drugs and alcohol are potentially a contributing factor for truck drivers, substance use is more common and more dangerous for drivers of passenger vehicles.

Similarly, Hanna et al. (2012) considered fatal crashes involving unlicensed young drivers (under age 19) in the U.S. using conditional and unconditional logistic modeling. Analysis was based on urbanicity (which categorizes all U.S. counties as urban, suburban or rural based on population and proximity to metropolitan areas) and the Townsend Index of Relative Material Deprivation (which serves as a proxy measure for socioeconomic status based on access to local goods, services, resources, and amenities). To allow for the simultaneous study of driver characteristics and region information, Adanu et al. (2017) used multilevel logistic modeling, which recognizes "the hierarchical structure in data and also provide[s] information to compute the amount of variability in the data attributable to each level of the hierarchy." They created a binary response variable which identified crashes as fatal or non-fatal. They used a two-level hierarchical logit model with driver characteristics at level 1 and regional information at level 2. In a sequential or hierarchical logistic regression model, explanatory variables can be added to the model step by step which permits examination of how the model changes with the addition of each set of variables. This approach would allow for the development of models at each level and understanding of the effects of these predictors on the response variable, at the driver level and regional level. Similarly, Chen et al. (2011) used multinomial logit models to examine the influence of drugs or alcohol on increasing the probability of injury or fatality for CMV drivers. Khorashadi et al. (2005) used a multinomial logit model to examine the effect of alcohol or drug use on rural road truck crashes. They concluded that the probability for severe/fatal injury increased 246 percent compared to crashes not involving alcohol or drugs.

Das et al. (2015) conducted an explanatory data analysis to develop a crash prediction model that estimated the likelihood of future crashes for at-fault drivers. They categorized the drivers into four types: not-at-fault prone drivers (involved in multiple crash but not responsible for), at-fault prone drivers (responsible for multiple crashes), not-at-fault non-prone drivers (involved in only one crash but not responsible for), and at-fault non-prone drivers (responsible for only one crash). An extensive data analysis was conducted to determine the association of these four driver categories with variables such as human-related factors, crash-related variables, roadway-related variables, environmental factors, and vehicle-related variable. The results of the data analysis emphasized the importance of understanding the behavior and other associated characteristics of drivers involved in multiple crashes (i.e., crash prone drivers). A logistic regression model was developed for crash-prone drivers, with the dependent variable being the fault status of the driver. The idea of categorizing the at-fault and not-at-fault drivers based on crash risk was a creative idea, however the model did not include all of them. The final model predicting the fault status was limited to only crash-prone drivers (i.e., drivers involved in more than one crash). To address this issue, a multinomial logistic regression modelling technique can be used which is an extension of binomial logistic regression. It allows for a dependent variable with more than two categories. In this case, the dependent variable can be split in four driver categories as defined by the researchers. Using multinomial logistic regression, the crash proneness (or any other categorical variables such as gender and educational attainment) can be added as a categorical explanatory variable. This will help to understand how categorical explanatory variables vary within the binary dependent variable. For example, this will help to determine how much more likely a crash-prone driver is to be at-fault than a non-crash prone driver.

Chandraratna et al. (2005) approached this scenario differently. They tried to predict the likelihood of a driver's involvement in a crash based on previous crash involvement. The dependent variable was whether or not the driver had a previous crash involvement observed during the study period. They used the fault

status of the driver as one of the independent variables. Results demonstrated that the drivers who had previous at-fault crashes were more likely to be involved in additional crashes than the remaining drivers. However, in this case a driver with one previous crash was considered riskier than a driver involved in five (for instance) previous crashes.

Other methods such as spatial analysis have also been used in crash analysis using socioeconomic factors. Brown (2016) considered the residential locations of at-risk drivers (drivers reported as contributing to fatal crashes) and the demographic characteristics associated with those residential locations at the Census Block Group level. Socioeconomic variables for higher risk block groups (more than 8 at-risk drivers per 1,000 driving population) were compared to those of lower risk groups to determine trends. This study used a cluster analysis to uncover hot spots of high or low risk areas that can be targeted for specific safety programs. Of note here, is the fact that this study examined demographic characteristics tied to the driver's home location instead of the commonly used method of socioeconomic characteristics tied to the crash location. Kocatepe et al. (2017) used hotspots to investigate the exposure of different age groups to severe injury crashes in the Tampa Bay region. The severity-weighted crash hotspots were identified using the Getis-Ord Gi method, weighted by the number of severely injured occupants involved in each crash. The study examined the proximity of residents in different age groups (17 and younger, 18 to 21, 22 to 64, and 65 and older) to severity-weighted crash hotspots. The results revealed that age, ethnicity, education, poverty level, and vehicle ownership impacted crash injury exposure.

A less defined but widely used method for this type of research simply involves separating crash or socioeconomic data into groups and comparing them with descriptive statistics. Abdalla et al. (1997) studied the effect of driver social circumstance on crash occurrence and casualty by linking crash records and census data in the Lothian Region, Scotland. The research showed a correlation between fatal crashes and a driver's distance from home. Socioeconomic variables were bundled into a Deprivation Index and postal codes were separated into the most affluent and most deprived to compare traffic casualties normalized by population. Similarly, Blatt et al. (1998) considered fatal crashes occurring in rural areas, with a focus on the residential location of the driver. Five years of crash data from FARS were linked with driver home zip code and other factors, including driver age, gender and blood alcohol concentration. Five levels of population density were identified for classifying each driver's residence location, including rural, small town, second city, suburban, and urban; other driver characteristics were divided into social clusters (e.g., age groups). Using geodemographic analysis, the percentage of drivers in fatal crashes in each social cluster was compared to the base population of that social cluster. In additional research involving traffic fatalities, Maciag (2014) investigated the differences in demographics of census tracts in relation to pedestrian fatalities in that tract. Census tracts were broken into categories by income and poverty to allow for a direct comparison of pedestrian fatalities.

2.4 Summary

In conclusion, the socioeconomic factors that seem most relevant to crash occurrence investigation are income, education level, poverty level, employment, driver age, and the rurality of an area. Education and income are typically negatively correlated with crash response, poverty is positively correlated, while employment varies across studies. Young drivers, and areas with a high proportion of young drivers, tend to be involved in a higher proportion of crashes and fatalities than other age cohorts, and in general, crashes in more rural areas are more fatal.

More research is merited to determine the role of alcohol and drugs in CMV driver crashes. Ideally, an analysis should be done to analyze the frequency of substance use among drivers in all truck collisions rather than only fatal collisions, but drug testing is neither common nor consistent for non-fatal crashes. Additionally, more research is needed to understand the truck drivers' fault in crashes involving drugs or alcohol, as the presence of substance seems to change severity of crashes.

Past research has shown a relationship between crash involvement and age. Aguero-Valverde et al. (2006) concluded that young drivers (under 25) and older drivers (over 65) face higher crash risks, and most of the previous literature has demonstrated a positive relationship between young drivers and crashes or fatalities. Several studies of older drivers identified their increased crash involvement and demonstrated the risk factors they create for themselves and others. Studies have also noted that young and old drivers have a positive relationship between crash involvement, indicating their higher propensity to be the at-fault driver in a crash. The current study further examines these trends to determine whether they hold for the Kentucky drivers.

Driver age, gender, and marital status (separated or widowed) have also been identified as good predictors of crash occurrence. In Kentucky, 55 percent of the drivers who were involved in collisions during 2016 (where the gender was listed) were male while 45 percent were female (KSP 2016). In fatal collisions, 74 percent of the drivers were male and 26 percent were female. Similar trends have been observed over the years and there might be crucial relationships between gender and crash occurrence (or crash severity) as it would be influenced by socioeconomic factors in Kentucky. In Alabama, Zephaniah et al. (2018) showed that DUI crashes are related to male employment and female education attainment. They considered several joint relationships (such as percent of males and females in different age groups, percent of population with different marital status, percent of population with different housing characteristics) in their spatial econometric analysis. However, only few were included in the final model. The percentage of drivers divorced and separated was considered in the preliminary analysis, however, it was not included in the final model due to multicollinearity. The current study investigates these interactions to determine whether they have an influence on crashes.

Previous crash records and citations are good predictors of crash occurrence. Though a couple of researchers attempted to include crash history/citation in their analysis, its relationship with crash occurrence was not defined completely. Das et al. (2015) investigated crash-prone drivers (with multiple crash records) to define their likelihood of being at-fault in a future crash, while Chandraratna (2005) tried to predict the likelihood of a driver being involved in a crash based on previous crash involvement. The former did not consider drivers with a single crash involvement, leaving room for future research. The latter used previous crash involvement as the dependent variable for predicting the likelihood of a driver with previous crash involvement being involved in a future crash. However, in this case a driver with one previous crash was considered as risky as the driver with five (for instance) previous crashes.

Apart from socioeconomic and demographic factors, potential crash-related variables and driver behavior are considered in this study. Few studies have attempted to relate human variables (e.g., restraint use, alcohol, ejection from vehicle, seating position) and vehicle variables (vehicle make, crash type, manner of leaving scene and most harmful event). The current study investigates these variables as they relate to the socioeconomic factors of the driver zip code and attempts to define possible relationships.

To investigate the role of these factors on crash occurrence, many different methods have been used, and while all of the considered methods are valid, there is still a wide range of analytical practices for relating socioeconomic characteristics with crash data. Many forms of regression techniques have been applied, as well as spatial statistics, clustering, and comparative grouping. The main objective of the current research is to identify factors that could potentially predict the fault status of a driver using the socioeconomic and demographic characteristics of their residence zip code. In other words, the response variable is the at-fault status of the driver, which is categorical in nature. In this case, logistic regression is the most appropriate and widely used method due to the categorical nature of the dependent variable. This modelling technique is beneficial when effects of more than one explanatory variable are examined. Binomial multiple logistic regression is used to estimate the probability of the driver's fault status based on multiple independent variables. This technique is applied for both CMV and automobile crash data to understand the influence of socioeconomic factors of the driver's residence zip code on crash occurrence. These types of models can

also be used to predict the probability that an observation falls into one of the categories of the dependent variable (i.e., at-fault determination). The log-odds for the response variable is a linear combination of explanatory variables which can be used to calculate the probability that the event occurs. Using the final model, specific groups of drivers vulnerable to crash risk could be identified, and then targeted for safety improvement programs in the future.

3. Research Methodology

The main objective of this project is to establish the relationship between crash occurrence and socioeconomic factors associated with the residence of the at-fault driver for both passenger car vehicles and CMVs. The analysis also considers other associated factors identified in the literature, including crash-related factors, driver characteristics, and the socioeconomic and demographic characteristics of the at-fault driver's residence zip code. The final model lets decision makers identify driver groups that need attention. The following section is a detailed description on the data and methodology used for the analysis.

3.1 Socioeconomic Descriptor Factors

The literature review identified several prominent factors that are relevant to and could explain crash occurrence — income, education level, poverty level, employment level, driver age, and the rurality of an area. Preliminary analysis showed typical correlations of these variables with crash occurrence; however, that analysis considered the crash data only for at-fault drivers (Cambron et al. 2018). These variables will also be evaluated in this study to address crash exposure in a more systematic manner and investigate how crash exposure could affect the association between these variables and crash occurrence.

Aguero-Valverde et al. (2006) concluded that age cohorts below 25 and over 65 have a positive association with crash risk and most of the previous research has shown a positive association between younger drivers and crashes or fatalities. Several studies on older drivers also identified their potential for increased crash involvement and the risk factors associated with them. This study also investigates these age groups in light of socioeconomic factors by grouping of drivers into age groups.

In addition to age, the literature review identified gender and marital status (separated or widowed) of the driver as good predictors of crash occurrence. Cambron et al. (2018) considered the percentage of drivers divorced and separated in their preliminary analysis, however, this was not included in the final model due to multicollinearity. The current study will investigate these interactions to determine whether they influence crashes, since the proposed approach considers crash exposure as well.

Previous research showed a well-defined relationship between level of education and crashes. The percentage of people with different education levels and their relationship linked with gender are also significant descriptors of crash propensity (Zephaniah et al. 2018). Further, race of the driver has also been identified as a factor associated with crash occurrence (Adanu et al. 2017). However, the research on the relationship between race and crashes is sparse. The current study will also evaluate the influence of major races (e.g., White, Black, American Indian, Asian and other races) on crash occurrence.

The negative correlation between income and poverty level and crashes has been previously established. These variables have an underlying relationship with rurality, education as well as employment. It is more likely that people with better education have better employment and higher income. These people tend to live in urban areas with better housing facilities. Therefore, it is expected that the housing characteristics of zip codes would also be a significant predictor of crash involvement.

The relationship between crash occurrence and previous crash records and citations is widely established. Depending on the availability of data, this information would be utilized as a predictive variable. This analysis is deemed appropriate, since the current study will evaluate prior driver history while considering the socioeconomic and demographic characteristics of at-fault driver residence.

Research has also shown that truck drivers are at risk for alcohol and drug use, both during their lifetimes and on the job, and that substance use seems to positively correlate with both truck driver injury severity and crash occurrence. For multivehicle truck crashes involving alcohol or drugs, the risk of serious injury

is twice as likely for the driver of the passenger vehicle compared to the truck driver. This study considers single and multi-unit crashes involving CMVs by looking at the socioeconomic factors of the at-fault driver's home residence.

3.2 Variable Selection Methods

Many socioeconomic variables need to be tested against driver at-fault status. It is tedious and time consuming to test all the possible combination of variables, to develop the best model with the most appropriate variables. As a first step towards variable selection and to better understand how socioeconomic variables could relate to driver at-fault status, two statistical analyses were conducted: correlation testing and recursive partitioning analysis.

3.2.1 Correlation Test

A correlation test is used to investigate the relationship between two variables. A point-biserial correlation coefficient measures the strength of association between a continuous variable and a binary variable (Laerd 2018). It is a special case of the Pearson's product-moment correlation coefficient (or Pearson correlation coefficient) which is applied when the correlation test is conducted for a binary variable. It measures the strength of association of two variables using a single measure, called a correlation coefficient (r), which ranges from -1 to +1. A coefficient value equal to -1 indicates a perfect negative association, a value of +1 indicates a perfect positive association, and a value of 0 indicates no association at all. A value greater than 0 indicates a positive association (i.e., as the value of one variable increases the value of the other variable also increases). A value less than 0 indicates a negative association (i.e., as the value of one variable increases the value of the other variable decreases). This test also calculates a p-value which represents the significance of the association between the two variables. This p-value is typically similar to a t-test output.

3.2.2 Recursive Partitioning Analysis

Recursive partitioning analysis is a statistical algorithm used for predictive modelling in statistics and machine learning (PennState n.d.). It attempts to correctly classify data along a decision tree, by splitting them into subgroups based on the variables at hand. It is an iterative process that builds a decision tree by sorting the independent variables down the tree based on how accurately they predict the target variables. The process continues until no more useful splits can be found. This method examines all the variables in the dataset to find the one that gives the best classification or prediction by splitting the data into subgroups. It helps in understanding the importance of the variables that should be considered in the modeling. The purpose of using this approach is to obtain a set of variables that can be used for logistic regression and modeling of at-fault probability.

The tests are conducted on both CMV and automobile data to aid the modelling process. Test results and their interpretations are described in the next section.

3.2.3 Identifying Interactions

Interactions offer a better understanding of the relationship among explanatory variables in a model. Inclusion of interaction terms, in addition to the main effects, is preferred for better mathematical stability of the model (Frost 2019). Two (or more) independent variables interact if the effect of one of the variables varies depending on the other variable(s). As noted above, there are several potential interactions among the socioeconomic variables that might influence crash occurrence. It is tedious and time consuming to test all the combinations of variables that can potentially form an interaction and for this reason, many previous analyses have not attempted to explore interactions. In some cases, interaction terms are identified based on prior knowledge and they are screened one by one. This research attempts to search optimal model containing interactions using an algorithm developed by the Department of Statistics at the University of Kentucky (FSA 2018). A tool called 'Shiny' uses a Feasible Solution Algorithm (FSA) for finding interactions. The algorithm allows for fixed, specified explanatory variables in the model and the addition of a feasibly best interaction (Lambert et al. 2018). It lets one formulate new or to improve upon existing

models. Several criterion functions (such as R^2 and adjusted R^2, interaction p-values, Akaike's Information Criterion and the Bayesian Information Criterion) are evaluated to examine model quality. FSA allows higher order interactions, however, this study is limited to two-way interactions.

Based on the results from variable selection methods, several combinations of explanatory variables are tested in the Shiny application to find the best solution. The results of the test are presented in the next section.

3.3 Crash Exposure – Quasi-Induced Exposure Technique

It is important to consider crash exposure when attempting to identify contributing factors to a crash. Crash databases do not contain information on driver exposure. Typically, vehicle miles travelled, number of licensed drivers, registered vehicles and other similar exogenous factors have been used to define exposure. With these conventional metrics, the exposure proportion of the driving population may vary depending on other factors such as time of day, driver gender or age, and road type. This has raised questions about the reliability and applicability of these exposure metrics when examining safety issues as they pertain to more specific groups of drivers or conditions, since the denominator in the ratio of crash occurrence for such subgroups and conditions cannot be obtained. The quasi-induced exposure technique developed by Carr (1969) overcomes this problem. The approach assumes that not-at-fault drivers represent the total population in question and the crash rate measure of exposure is developed in terms of the relative accident involvement ratio (RAIR), which is the ratio of the percentage of at-fault drivers to the percentage of not-at-fault drivers from the same subgroup.

This ratio is defined in Equation 1:

$$\text{RAIR} = \frac{\text{proportion of at-fault drivers}}{\text{proportion of not-at-fault drivers}} \quad (1)$$

Chandraratna and Stamatiadis (2009) examined the validity of this assumption using two samples of not-at-fault driver data: one with not-at-fault drivers selected from the first two vehicles in a multi-vehicle crash and a second that included not-at-fault drivers (excluding the first two drivers) from multi-vehicle crashes with more than two vehicles involved. They concluded that the two samples were statistically the same and stated that "estimating relative crash propensities for any given driver type by using the quasi-induced exposure approach will yield reasonable estimates of exposure."

3.4 Statistical Modeling

As discussed in the literature review, logistic regression is the most appropriate and widely used method when the dependent variable is categorical in nature. This modeling technique is beneficial when effects of more than one explanatory variable influence an outcome (Das et al. 2015). The independent variables can be discrete and/or continuous. In linear regression, the expected values of the response variable are modeled based on combination of predictor values while in logistic regression, probability or odds of the response taking a particular value, is modeled based on combination of predictor values.

Mathematically, a logistic regression estimates a multiple linear regression function defined as:

$$y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (2)$$

where y is the dependent variable, X's are the explanatory variables, a is the intercept and b's are the coefficients of the explanatory variables. In this case, the left-hand side of the equation could result in negative values or values greater than 1, while y (the dependent variable) is categorical (i.e., y has a value of 0 or 1). This problem is solved by taking the logarithm of the odds of the response variable. Therefore,

the logistic regression defines the log-odds for the response variable as a linear combination of explanatory variables.

The logarithm of odds, which is otherwise called log-odds or the logit function, is defined as:

$$\text{logit (p)} = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characterestic}\right) \quad (3)$$

where p is the probability of presence of the characteristic of interest.

In the context of the crash analysis used here:

$$\text{logit (p)} = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{\text{probability of being at-fault}}{\text{probability of not being at-fault}}\right) \quad (4)$$

where p is the probability of a driver being at-fault. Here, the ratio between the probability of at-fault drivers to the probability of not-at-fault drivers is equivalent to the RAIR, which is the driver exposure measure in the quasi-induced exposure technique.

By taking the log-odds of the response variable, equation 2 changes to:

$$\text{logit (p)} = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

i.e.,
$$\ln\left(\frac{p}{1-p}\right) = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (5)$$

After taking the anti-logarithm of Equation 5 and replacing the regression equation with $f(X)$, the equation for the probability of the characteristics of interest is expressed as a function of the regression equation:

$$p = \frac{e^{f(X)}}{1 + e^{f(X)}} \quad (6)$$

On further mathematical manipulation, equation 6 takes its final form,

$$p = \frac{1}{1 + e^{-f(X)}} \quad (7)$$

$$\forall \quad f(X) = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Where $f(X)$ is the regression model, X_i are the explanatory variables, a is the intercept and b_i are the coefficients estimated using the maximum likelihood method.

3.4.1 Assumptions

The data must meet different assumptions of logistic regression to produce valid results (Laerd 2018). In practice, the data may fail to meet certain assumptions, however, there are solutions to overcome this. If a violation of an assumption is not correctable, binomial logistic regression is not recommended for the dataset. Various tests are conducted to ensure that the data properties satisfy the assumptions of logistic regression. The major assumptions of logistic regression and test results are described below.

Assumption 1 — The dependent variable should be measured in dichotomous scale (i.e., binary). Examples of dichotomous variables are gender (two groups: males and females), presence of heart disease (two groups: yes and no), and so forth. Here, the dependent variable is 'Fault status' of a driver involved in a crash, which has two possible groups: at-fault or not-at-fault. The variable is coded using 0's and 1's to represent the at-fault and not-at-fault driver groups, respectively.

Assumption 2 — There are one or more independent variables which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable).
Crash data used for this project have a combination of continuous and categorical variables which are the potential independent variables in the regression model.

Assumption 3 — Data should have independent observations and the dependent variable should have mutually exclusive and exhaustive categories.
Crashes are independent of each other and the occurrence of one does not affect the probability of another occurring. An event which is both collectively exhaustive and mutually exclusive can take one only value at a given time. In this study, the dependent variable is the fault status of a driver; hence a driver involved in a crash must be either at-fault or not-at-fault.

Assumption 4 — The dataset used for logistic regression typically requires a large sample size.
Logistic regression assumes linearity of independent variables and log odds. Although this analysis does not require that dependent and independent variables be linearly related, it requires independent variables be linearly related to the log odds. Crash data used for this research forms a large dataset which is adequate for logistic regression. Also, the potential socioeconomic independent variables were tested for linearity to the log odds, satisfying this assumption.

All four assumptions were tested and satisfied. Therefore, logistic regression is recommended for the dataset used here.

3.4.2 Relative Accident Involvement Ratio
Binary logistic regression is used in this research to develop a regression model to predict the fault status of the driver based on different socioeconomic and demographic variables. Equation 7 allows for the estimation of the likelihood of a driver belonging to a particular zip code (with specific socioeconomic and demographic factors) being the at-fault driver in a crash. Here, p is the probability of a driver being at-fault in a crash, while considering as exposure, drivers with the same characteristics who were not at-fault in a crash. Equation 7 is analogous to the RAIR used in the quasi-induced exposure methodology measures crash propensity (as discussed in the previous section).

Considering the probability of a driver at-fault being calculated as *p*, the RAIR of a driver group can be calculated using Equation 8.

$$\text{RAIR (at-fault)} = \frac{\text{p}}{1-\text{p}} \tag{8}$$

The following example demonstrates the use and interpretation of RAIR. Stamatiadis and Puccini (1999) indicated that in the U.S. Southeast males have a higher rate for fatal crashes — 78 percent for single-vehicle and 70 percent for multivehicle crashes. Looking at exposure data, males represented 73 percent of the driving population involved in multivehicle crashes. When they analyzed the involvement ratios by gender, they concluded that even though males are more likely to cause single-vehicle crashes, females are more likely to cause multivehicle crashes (Figure 1). This may be explained by the different levels of risk that each gender is willing to take.

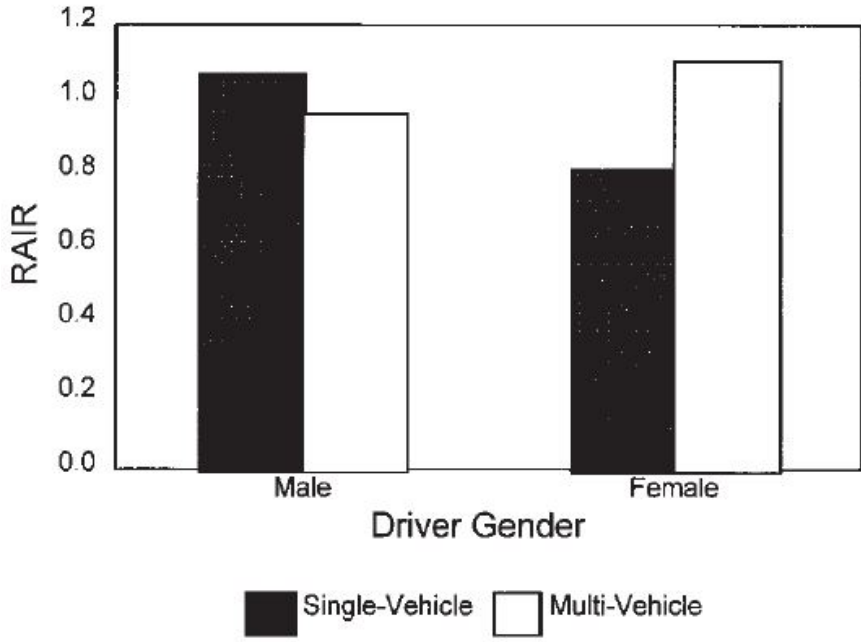


Figure 1 RAIR for Driver Gender

The quasi-induced exposure approach is used here to define driver exposure by assuming that the not-at-fault drivers represent the general population. In this study, the response variable is categorical (i.e., at-fault and not-at-fault driving status of the driver) and logistic regression is the most appropriate method to analyze this binary dependent variable.

Based on the probabilities developed using logistic regression, target groups/target areas with high crash propensity are identified for more detailed examination. This allows policymakers to focus their efforts to improve safety through targeted efforts and specific road safety campaigns.

3.4.3 Evaluation Criteria

Several models are developed for CMVs and automobiles based on qualitative and quantitative variable selection. These models have undergo several evaluations to produce the best possible model. The model evaluation criteria are explained below.

Likelihood Functions

The two likelihood functions used for the model evaluations are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). They are estimators of the relative quality of statistical models for a given dataset and criteria for model selection among a finite set of models. Models with the least likelihood function are preferred. One of the main drawbacks of these criteria is the possibility generating an increase in likelihood by adding more parameters, which may result in overfitting.

Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the performance measurements of a model. It is a probability curve plotted between the true positive rate (or sensitivity) and false positive rate (or 1-specificity) and represents the model's ability to distinguish between the two classes (i.e., driver at-fault status). The area under the curve (AUC) represents the degree or measure of separability between the two classes. An excellent model has AUC near 1, which means it has good measure of separability. A poor model has AUC closer to 0, which means it is reciprocating the result (i.e., it predicts 0's as 1's and 1's as 0's).

Training and Validation Method

In this method, the dataset is randomly divided into two parts — a training set and a validation set. The model is developed using the training set and the fitted models are used to predict the responses for the validation set. The percentage correctly predicted is calculated to evaluate the model's ability to represent the data. In general, the training set is larger than the validation set to ensure that the training set is a good representation of the overall dataset. Here, an 80:20 percent split is used to divide the dataset into training and validation sets.

These criteria are tested for each model developed to come up with the most appropriate one in each case.

3.5 Model Development Approach

Many socioeconomic variables need to be tested against driver at-fault status. To simplify the tedious and time-consuming process of testing all the possible combinations of variables, two statistical analyses were conducted: correlation and recursive partitioning analysis. These processes reduce the number of factors or predictors that need to be considered in the model and their results are used as a starting point for the logistic regression model development. Correlation procedures are used to investigate the relationship between the dependent variable and the socioeconomic variables. This test calculates a p-value that measures the significance of the association between the variables. Variables that are statistically significant provide a starting point for variable selection.

Since the dependent variable in this study is categorical, the Pearson coefficient may not be an appropriate measure to explain the relation between crash occurrence and the socioeconomic variables. Instead the recursive partitioning analysis may be more appropriate, which is another statistical technique used to understand the relationship between the potential predictor and dependent variables. It helps by generating a tree-like model that aids in variable selection when the dependent variable is categorical. This approach is used to obtain a set of variables that can be used in the logistic regression model for predicting the at-fault driver status. This method examines all the variables in the dataset to find the one that gives the best prediction by splitting the data into subgroups. This approach provides a relative importance among the variables being considered and indicates those variables that should be given priority for inclusion in the logistic regression.

The results from the two techniques are used for the statistical modeling. In addition to the variables identified through these analyses, other variables are considered and tested to finalize the model with the most appropriate set of predictors. For example, if the education variable 'Percent below high school graduate' is a descriptor of note in the recursive partitioning analysis, it is considered first in the modeling. However, the other education variables (such as 'percent with high school graduate' and 'percent with bachelor's degree), which are significantly related to the dependent variables, based on the correlation test, are also tested. Each variable from the socioeconomic categories is tested to identify the best representation of that category in predicting at-fault driver crash involvement. Multiple variables from the same category are not used in the same model to avoid complimentary effect. Several models are developed for single and two-unit crashes using this approach and their parameters are evaluated using the above-explained criteria for selecting the final model.

4. Data Collection and Preparation

4.1 Crash Data

Kentucky crash data, aggregated at the zip code level, are used to examine the characteristics of drivers involved in crashes. The crash data cover the 2013-2016 period and were collected from the Kentucky State Police (KSP) records; they include the 5-digit zip code of the driver residence. About 77 percent of the crashes that occurred during the four-year time period were two-unit crashes, 13.7 were single-unit crashes, and the remainder involved three or more vehicles. This research primarily focuses on single and two-unit crashes, which restricts the number of drivers involved to a maximum of two. The variables listed in Table 1 were extracted from the KSP database. Information on crash severity, manner of collision, roadway characteristics, vehicle type, weather condition, and lighting conditions, and are used in this analysis.

Table 1 List of Crash Record Variables

Variable Type	Variable
Crash	Master file number
	Year of collision
	Severity of crash (KABCO)
	Number of people injured
	Number of people killed
	Collision date & time
	Collision day week code
	Intersection crash indicator
	Number of units involved
	County code
	Crash location in lat\long
Vehicle	Unit number
	Unit type code
	Vehicle year
Roadway condition	Total number of lanes
	Roadway character code
	Roadway surface code
	Roadway condition code
	Weather code
	Light condition code
	Land use code
	Function class code
Person	Person number
	Person type code
	Zip code of driver residence
	Age at collision time
	Gender
	Human factors detected

This study does not consider information on passengers and pedestrians since driver at-fault status in a crash is key to the methodology.

The human factors coded for each driver are used to determine at-fault status. For each crash, the driver with a human factor code recoded by the police officer is considered to be the at-fault driver for the crash (Chandraratna and Stamatiadis 2009). In the crash database, multiple human factors are recorded (if any) for drivers involved in crashes. For example, if there are three human factors recorded for a driver involved in a two-vehicle crash, there will be three entries for that particular Master File Number (MFN, a unique number identifying each crash). After data processing in Python, the human factors recorded to the same driver are aligned to convert the multiple entries to a single entry. Age and gender of the driver are used as the factors to correlate the entries belonging to the same driver. The first human factor recorded is used to determine the driver's at-fault status. For each MFN, the driver with the first human factor coded as 'non-detected' is considered to be not at fault, while the driver with a human factor detected is treated as the at-fault driver. Crashes in which a human factor code is recorded for both or neither drivers are eliminated from the analysis. This selection criteria avoids multiple at-fault drivers for the same crash in two-unit crashes (Chandraratna and Stamatiadis 2009). In single-unit crashes, only drivers with a human factor coded are included in the dataset, and these drivers are coded as at-fault. As single-unit crashes have only one vehicle involved, there is no not-at-fault driver group involved these crashes. Therefore, the not-at-fault driver group from the two-unit crashes were included in this dataset to facilitate use of the quasi-induced exposure technique.

4.1.1 Commercial Vehicle Crash Data

To analyze the relationship between CMV drivers involved in crashes and their socioeconomic characteristics, the CMV crashes are extracted from the previously prepared dataset. Two datasets with single-unit and two-unit CMV crashes are prepared and linked to the socioeconomic data of the driver residence.

Kentucky crash data from 2013 to 2016 have 9,248 cases where a CMV driver was involved either as an at-fault or not-at-fault driver. Among those crashes, 8,584 involved a CMV and a passenger car, and 332 crashes involved two CMVs. Over the same time period, there were 2,955 single-unit CMV crashes. Single-unit crashes have only one vehicle involved, thus the driver is considered as the at-fault driver. To properly account for the crash involvement of drivers in single-unit crashes using the quasi-induced exposure technique, drivers without a human factor coded are excluded from the analysis (Stamatiadis and Deacon 1997). At the same time, the not-at-fault group from the two-unit crashes are included in the quasi-induced exposure analysis of single-unit crashes to account for driver exposure. The sample size of the not-at-fault group of drivers in the two-unit crashes is 4,798 and this is used to form the exposure for the single-unit CMV crashes.

The final CMV datasets also include drivers with ages between 15 and 90 years. These drivers are grouped into seven categories: 20 <, 20-24, 25-39, 40-64, 65-74, 75-84 and > 85. The distribution of age groups for single-unit and two-unit truck crashes are shown in Table 2. The data indicate a small number of drivers belong to the youngest and the oldest groups. Therefore, the first two and the last two groups are combined to allow for more significant categories of age groups.

Table 2 Commercial Vehicle Driver Age Distribution (2013-2016)

	Single-unit							
	Age Group							
Fault Status	<20	20-24	25-39	40-64	65-75	75-84	>84	Total
At-fault	33	210	951	1,556	174	30	1	2,955
	Two-unit							
	Age Group							
	<20	20-24	25-39	40-64	65-75	75-84	>84	Total
At-fault	59	318	1,333	2,381	280	72	7	4,450
Not-at-fault	36	222	1,393	2,881	231	33	2	4,798
Total	95	540	2,726	5,262	511	105	9	9,248

Fatigue and substance use are believed to be a major cause of CMV crashes. Therefore, it is important to examine whether the human factors recorded have any relationship with the socioeconomic factors. In this analysis, the dependent variable is the human factor code while the explanatory variables are the socioeconomic variables discussed previously. Based on the assumption that the driver with a human factor recorded is the at-fault driver, the analysis on human factors is conducted only using at-fault drivers.

Twenty-six human factors are recorded for CMV crashes. It is not statistically meaningful to use a dependent variable with 26 categories. Therefore, human factors are categorized into two groups: behavioral and non-behavioral. Human factors marked in bold in Table 3 are grouped together as behavioral factors while the remaining are grouped as non-behavioral factors. Hence, the dependent variable is categorical and binary logistic regression is used to test them against socioeconomic variables.

Table 3 Human Factor Code Frequency for Commercial Vehicle Crashes (2013-2016)

Human Factor	Single-unit	Two-unit
Alcohol involvement	23	8
Cell Phone	4	3
Disregard traffic control	10	71
Distraction	52	116
Drug involvement	8	3
Emotional	3	3
Exceeded stated speed limit	3	2
Failed to yield right of way	1	465
Fatigue	23	5
Fell asleep	41	7
Following too close	4	230
Improper backing	46	161
improper passing	1	34
Inattention	728	1728
Loss of consciousness/Fainted	17	3
Medication	3	1
Misjudge clearance	984	760
Not under proper control	505	326
Overcorrecting/oversteering	175	13
Physical disability	2	1
Sick	10	5
Too fast for conditions	74	42
Turning improperly	14	88
Weaving in traffic	2	1
Other	222	374

4.1.2 Automobile Crash Data

The 2013-2016 Kentucky crash data contain 239,722 two-unit crash pairs. At the same time, there are 119,592 single-unit crashes, while only 74,751 of them have a human factor recorded. As noted, the driver in single-unit crashes is considered the at-fault driver. The not-at-fault group from the two-unit crashes are included in in the quasi-induced exposure analysis of single-unit crashes to account for driver exposure. The sample size of the not-at-fault group of drivers in the two-unit crashes is almost 3.2 times larger than the at-fault group of single-unit crashes. To avoid the sample size disparity of not-at-fault group in the data, a random sample equivalent to 75,000 is drawn from the original not-at-fault group. This sample is used as the not-at-fault group of drivers in the single-unit crash data.

The data are processed using the above-explained procedure to develop the final dataset for single-unit and two-unit automobile crashes. The final dataset includes drivers with ages between 15 and 90 years. To analyze the RAIR of drivers in different age groups, ages are categorized into seven groups — <20, 20-24, 25-39, 40-64, 65-74, 75-84 and >85. Table 4 shows the distribution of age groups in the dataset.

Table 4 Automobile Driver Age Distribution (2013-2016)

	Single-unit							
	Age Group							
Fault Status	<20	20-24	25-39	40-64	65-75	75-84	>84	Total
At-fault	11,803	13,239	22,784	21,467	3,461	1,646	351	74,751
	Two-unit							
	Age Group							
	<20	20-24	25-39	40-64	65-75	75-84	>84	Total
At-fault	30,367	36,325	68,051	74,886	17,987	9,825	2,281	239,722
Not-at-fault	14,673	24,794	72,161	102,276	18,710	6,194	914	239,722
	45,040	61,119	140,212	177,162	36,697	16,019	3,195	479,444

The commercial and automobile crash datasets are prepared using the process explained above. Table 5 shows the number of crashes per unit group in each case.

Table 5 Sample Size of the Datasets

		Automobiles	Commercial Vehicle
Single-unit	At-fault	74,751	2,955
	Not-at-fault	75,000	4,798
	Total	149,751	7,753
Two-unit	A-fault	239,722	4,450
	Not-at-fault	239,722	4,798
	Total	479,444	9,248

4.2 Socioeconomic Data

Socioeconomic and demographic variables were collected from the American Community Survey (ACS). The ACS database has two sets of information significant for this research: People and Housing. The People category includes general information on the population (e.g., total population, race, marital status, age, gender, education, income, employment, poverty status) while the Housing category includes data on households (e.g., value of house, number of housing units, household size, household type) in a particular geographical area. The choice of variables is made in reference to the findings and suggestions of previous research and the initial analysis conducted as part of this effort. Table 6 lists the socioeconomic variables chosen for analysis. They are divided into six major categories — Race, Housing, Marital Status, Education, Income, and Other.

Table 6 List of Socioeconomic Variables

Category	Variable
Race	Percent white
	Percent black
	Percent American Indian
	Percent Asian
	Percent other races
Housing	Household units
	Household ownership total
	Owner occupied housing units
	Renter occupied housing units
	Median housing value
Marital Status	Percent now married
	Percent widowed
	Percent divorced
	Percent separated
	Percent never married
Education	Percent less than high school graduate
	Percent high school graduate
	Percent some college or associate degree
	Percent bachelor's degree or higher
	Percent graduate or professional degree
Income	Median individual income
	Mean individual income
	Household mean income
	Household median income
Other	Employment population ratio
	Percentage rural
	Unemployment rate
	Percent below poverty level
	Total population

According to 2016 population estimate, 85 percent Kentucky's population is comprised of White Americans, followed by 8.3 percent of Black Americans (KPH 2017). Adanu et al. (2017) indicated race is a factor associated with crash occurrence. However, the research on association of races with crashes is sparse. This research tests the relationship between race and crash occurrence. Therefore, proportion of major races (White, Black, Indian, Asian, and Others) were extracted from population estimates of the ACS. The other races include the sum of proportion of population belonging to races such as Native Hawaiian and other Pacific Islander alone, Two or More Races, and Hispanic or Latino. The information on all the races is included in this dataset for further investigation.

Housing is another category of variables, and it is a well-established predictor of crash involvement. Housing density is most frequently considered as a surrogate for level of rurality for a state. Noland and Quddus (2004) and Hasselberg et al. (2005) offered an explanation for the relationship between housing

and unsafe traffic conditions. Lower income people tend to live in rural areas where cost of living and housing are cheaper. These places are less likely to have adequate infrastructure and safe traffic conditions. Therefore, the number of household units and median housing value are considered and will be included in the analysis as they are surrogate indicators of rurality. It is also noted that the areas with high rental housing percentages have exhibited lower DUI crash rates (Zephaniah et al. 2018). It is important to examine the potential effect of different housing ownership levels on crash occurrence. Therefore, data on housing characteristics (rental/owned) are also included in this analysis.

Marital status is expected to have a significant relationship with crash occurrence; however, this association has not been adequately established. Factor et al. (2008) provided evidence that separated and widowed drivers are 50 percent more likely to be involved in a crash than married drivers. Stressful life events may inhibit safe decision-making, resulting in increased risk of causing a crash. Information on the proportion of the population currently married, previously married (widowed, separated, and divorced) and never married are included in the dataset for further investigation.

Several researchers have investigated the correlation between educational level of drivers and their involvement in crashes to discover patterns that can prevent or decrease crashes. It has been noted that people with low educational attainment account for the highest mortality rate (Hasselberg et al. 2005 and Zephaniah et al. 2018). Cook et al. (2005) discussed a positive relationship between female education achievement and crash involvement. This joint relationship between gender and educational attainment is further tested here as well to examine any possible relationships for Kentucky.

Income is another relevant predictor for crash-related analysis. Personal and household income are cited as significant explanatory variables to crashes; however, personal income is more widely used as the socioeconomic variable representing income (Stamatiadis and Puccini 1999; Noland and Quddus 2004; Zephaniah et al. 2018; Lee et al. 2014). This research considers both household and personal incomes to identify the one most representative for the Kentucky drivers in relation to crash prediction. Therefore, different mathematical representations (mean and median) of both individual and household income were extracted from the Census database.

Other well-established predictors of crashes include employment rate, poverty level, and rurality. These variables are correlated with income, housing, and education. Their interdependency is also explored in this analysis.

Information on the above discussed variables was obtained from a five-year estimate of 2016 ACS data at the zip code level. The data for the demographic and socioeconomic descriptors are joined at the zip code level and then merged to the data of crash-related variables matching the residence zip code of the driver. Python tools are used to prepare the final dataset. The variables in the final dataset are tested with the dependent variable (at-fault status of the driver) to understand their relationship with each other. Variables indicating correlation with the dependent variable in the initial correlation analysis are used for the final regression modelling. Multiple variables from same socioeconomic category are not used in the same model to avoid dependency. For example, percent white and percent non-white are complementary, and it would produce ambiguous results if they were considered in the same regression model.

To examine the effect of the socioeconomic characteristics of the driver's residency zip code on CMVs, a two-step analysis is performed. First, models are developed that correlate possible socioeconomic variables with crash occurrence and driver at-fault status for CMVs. Second, these models are compared to the models developed for automobiles to determine whether CMV drivers behave in a different manner than drivers of other vehicles. This process not only identifies potential socioeconomic variables that would be of interest for improving the safety of CMV drivers, but it also determines whether they are different than other drivers in that respect.

5. Statistical Modeling

This research primarily focusses on two-unit and single-unit-crashes of CMVs and automobiles. The objective of this effort is to identify the strongest socioeconomic variables that could be used as predictors for both CMV and automobile crashes, and compare them with each other, to identify how the CMV crashes differ from automobile crashes. The final dataset includes several socioeconomic variables as described in the previous section. In addition to the most discussed descriptors (e.g., income, education, employment) by previous researchers, many other variables (e.g., race, housing characteristics, marital status) are also included in the analysis. Several preliminary tests are conducted to make the appropriate choice of variables for the modelling.

Correlation tests are used to first examine the significance of each socioeconomic variable in predicting the dependent variable. The variables that have a statistically significant relationship with the dependent variables are narrowed down for a starting point in variable selection. Recursive partitioning is then used to better understand the association between the potential predictor and dependent variables. This step helps clarify the importance of the variables that should be considered in the modelling. These two tests are conducted on the commercial as well as the automobile crash data, and then used input in variable selection for the modeling. Based on their results, binary logistic regression models predicting crash occurrence are developed for each dataset.

5.1 Commercial Vehicles

To analyze the relationship between CMV drivers involved in crashes and their socioeconomic characteristics, datasets are prepared following the steps explained in the methodology. Two datasets with single-unit and two-unit truck related crashes are prepared and linked to the socioeconomic data of the driver residence zip code. Correlation tests and recursive partitioning analysis are conducted on the dataset to develop a preliminary variable selection. Test results are discussed in the next section. Based on the results of the variable selection process, several models are tested and their model parameters compared to recommend the most appropriate model for determining the probability of a driver being at fault.

5.1.1 Variable Selection

Correlation Test

A correlation matrix is developed to identify variables associated with at-fault status. Pearson correlation coefficients are computed for each variable which represent its relationship with the dependent variable. Variables that are statistically significant at 95 percent are marked in green in Table 7.

The correlation test conducted here for each socioeconomic characteristic identifies those that are significantly related to the driver's at-fault status. P-values less than 0.05 are considered to be significantly correlated with the at-fault status at the 95 percent level in a two-tailed test. The arithmetic sign of Pearson coefficient indicates the direction of the relationship between the socioeconomic variable and the indicator of crash occurrence. For example, the income variables are positively correlated with the at-fault status, which means that as the income of the driver residence zip code increases, the likelihood of causing a crash increases. The explanation of the results shown in Table 7 is discussed below. These coefficients are compared with the correlation test results of automobiles to further determine whether there are differences between the two vehicle groups with respect to the ability of the socioeconomic variables to predict at-fault probability.

Among the five categories of race, the proportion of white, black and Asians is significantly correlated with two-unit truck-related crashes. No predominant relationship is observed between races and single-unit truck crash occurrence. However, proportion of white and black seems to be potential descriptors of crash occurrence for two-unit truck related crashes. Even though, these categories are not significant for single-

unit truck crashes, these variables are considered in the statistical modelling to examine whether they show any significance when considered along with other variables.

Table 7 Correlation Analysis Results for Commercial Vehicle Crashes

Category	Variable	Single-Unit		Two-unit	
		Correlation Coefficient	P-value	Correlation Coefficient	P-value
Race	Percent white	-0.007	0.563	-.035**	0.001
	Percent black	0.008	0.461	.031**	0.003
	Percent American Indian	0.009	0.413	-0.002	0.817
	Percent Asian	-0.002	0.860	.049**	0.000
	Percent other races	-0.006	0.608	0.009	0.383
Housing	Household units	-0.014	0.608	.034**	0.001
	Household ownership total	-0.018	0.122	.034**	0.001
	Owner occupied housing units	-.032**	0.005	.030**	0.004
	Renter occupied housing units	0.004	0.722	.035**	0.001
	Median housing value	-.041**	0.000	.043**	0.000
Marital Status	Percent now married	-0.018	0.104	-.022*	0.034
	Percent widowed	.030**	0.008	-0.020	0.059
	Percent divorced	-0.002	0.869	-0.003	0.773
	Percent separated	0.012	0.282	-0.004	0.709
	Percent never married	0.011	0.319	.032**	0.002
Education	Percent less than high school graduate	.038**	0.001	-.031**	0.003
	Percent high school graduate	-0.020	0.086	-.046**	0.000
	Percent some college or associate degree	-0.011	0.334	0.019	0.071
	Percent bachelor's degree or higher	-0.011	0.315	.045**	0.000
	Percent graduate or professional degree	0.004	0.692	.041**	0.000
Income	Median individual income	-.050**	0.000	.033**	0.001
	Mean individual income	-.034**	0.003	.031**	0.003
	Household mean income	0.000	0.439	.026*	0.012
	Household median income	-.049**	0.000	.022*	0.037
Other	Employment population ratio	-.061**	0.000	.039**	0.000
	Percentage rural	0.019	0.088	-.045**	0.000
	Unemployment rate	.027*	0.016	-0.003	0.754
	Percent below poverty level	.054**	0.000	-0.010	0.336
	Total population	-0.021	0.060	.032**	0.002

All of the housing variables are statistically associated with two-unit truck related crashes, however, housing density is not related to single-unit truck crashes. As discussed previously, housing value is another factor which is related to rurality. This could be related to household income, as families with high incomes tend to live in areas with high housing values. Housing ownership characteristics (rental/owned) are also

correlated with two-unit truck related crashes, while rented house density is not related to the occurrence of single-unit truck crashes. These relationships are further investigated in the final modeling.

Marital status has no substantial effects on two-unit and single-unit truck crashes. Percent now married and percent never married are significantly related to the occurrence of two-unit truck crashes, whereas percent widowed is the only significant representation of marital status in single-unit truck crashes. A detailed investigation on the effect on marital status on the occurrence of truck-related crashes is pursued in the next level of analysis. Furthermore, education seems to be a potential descriptor of two-unit truck related crashes, while it does not have a significant relationship with single-unit truck crashes. Again, its relationship with truck crashes requires more investigation.

Individual, as well as household, income show significant relationship with the at-fault status of the truck driver. Prior research (Stamatiadis and Puccini 1999; Lee et al. 2014) has demonstrated household income is a better predictor of crash occurrence. Analysis on trucks examines the various income categories and determines the most appropriate one for inclusion in the final model for predicting occurrence of truck-related crashes. Also, other variables, such as rurality, poverty level, and employment by population ratio, that have well established relationships with truck-related crashes, may be also correlated with income and educational level and their interactions is examined.

Based on the outcomes observed in the preliminary analysis, logistic regression models describing fault status of single and two-unit truck related crashes are developed. The process of model development is described in the next section.

Recursive Partitioning Analysis

Recursive partitioning analysis is performed on the single-unit truck dataset to develop the tree-like model assisting variable selection when the dependent variable is categorical. The classification tree identified education as one of the most important factors influencing single-unit truck crashes. Meanwhile, income, age group, percent white, and percent widowed were other variables of the classification model. Based on the input from the classification model and the correlation test, different variables are tested to develop the most suitable model representing single-unit truck crashes. Note that multiple variables from the same category were not used in the same model to avoid complementary and dependency effects. For example, percent white and percent non-white are complementary, and it would show misleading results if both of them are considered in the same model.

In the two-unit truck dataset, the tree-like model identified employment by population, income, percent never married, percent high school and college, and household units as the best descriptors. Based on the input from the classification model and the correlation test, different variables were tested to select the most suitable for representing two-unit unit truck related crashes. Again, multiple variables from the same category are not used in the same model to avoid complementary effects.

5.1.2 Regression Models

Based on the results from the variable selection process, several models are tested and their model parameters (such as AIC, BIC, AUC and Log-likelihood) compared to choose the final model. The training and validation approach is also used for the model development to estimate accuracy in predicting the dependent variable. Final models of single-unit and two-unit CMV crashes are described below.

Single-Unit Commercial Vehicle Model

The models developed for single-unit CMVs are given in Table 8. As noted in Table 2, drivers were grouped in age groups and initially younger (under 25) older (over 75) were consolidated into single groups due to low sample sizes. To further improve the model's predictive power, additional groupings of driver age groups are tested. Those shown in Table 8 resulted in a model with better predictive power.

Table 8 Models for Single-Unit Commercial Vehicle Crashes

Variable	Model 1			Model 2		
	Estimate	Wald	p-value	Estimate	Wald	p-value
>25		10.000	0.019		9.957	0.019
25-39	-0.353	5.033	0.025	-0.331	4.179	0.041
40-64	-0.442	8.589	0.003	-0.457	8.671	0.003
>65	-0.222	1.205	0.272	-0.280	1.814	0.178
Median housing value	-3.34E-05	312.546	0.000	-3.39E-05	269.171	0.000
Percent with high school and some college education	-0.463	1776.998	0.000	-0.483	1786.010	0.000
Household median income	6.84E-05	127.382	0.000	3.11E-04	226.228	0.000
Employment by population ratio	0.147	429.114	0.000	0.313	439.888	0.000
Employment by population ratio × Household median income				-4.21E-06	150.897	0.000
Constant	26.396	1653.493	0.000	18.549	452.782	0.000

Model 1 is the simplest model for estimating at-fault driver propensity based on socioeconomic factors for single-unit CMV crashes. This model defines probability of fault as a function of age group, household median income, median housing value, an education indicator, and employment by population ratio. All variables in the model are significant at the 95 percent confidence level.

Several variables describe education levels, such as, 'percentage below high school graduate' and 'percent high school graduate,' are examined to determine which results in the strongest model. 'Percent of high school and some college education' is the one that best represented education in the model. Models tested with other categories of education dropped the AUCs to 0.50. The estimates of Model 1 imply that the truck drivers with high school education and/or some college education are likely to cause fewer crashes. CMVs have a significant relationship with level of education of the drivers in a zip code. The Wald chi-square for this education category is significant and its inclusion overall improved the model prediction power.

Followed by education, employment by population ratio, median housing value and household median income are the other important socioeconomic descriptors in the model. Recursive partitioning analysis recommended both individual and household income as good predictors. However, the model using median household income is the best indicator of a driver's economic condition and agreed with the results from prior research. Age groups is another predictor variable in the model. Based on the results, younger and older groups of CMV drivers are more likely to cause a crash. Percent white and percent widowed are also recommended from the classification model (developed in the recursive partitioning analysis), however, their p-values are high and the variables not significant in the logistic model. Any efforts to force them into the model result in lower overall model parameters including AUCs and AIC/BIC.

Interactions are also tested using the FSA. A two-way interaction between income and employment by population ratio is identified. Model 2 in Table 8 incorporates this interaction. Table 9 compares performance of the two models.

Table 9 Statistics of Single-Unit Commercial Vehicle Models

Criterion	Model 1	Model 2
Akaike information criterion (AIC)	4721.6	4484.52
Bayesian information criterion (BIC)	4777.18	4547.04
Area under Curve (AUC)	0.943	0.948
Percentage Correctly predicted	88.6	88.9
Chi-Square (DF)	5523.789 (7)	5762.876 (8)
Prob>ChiSq	<0.0001	<0.001

The data in Table 9 indicate the model improves when the interaction term is included based on the values of AIC and BIC. However, this is improvement is not significantly large. The AIC and BIC of Model 2 are lower, implying improved model predictability. However, the AUC, which represents the model's ability to distinguish between the two categories of the predictor variable, remain more or less same. Therefore, to maintain simplicity in the mathematical framework, Model 1 (Table 8) is preferred.

The odds ratios obtained from the statistical analysis, along with the calculated probability to be the at-fault driver and RAIR, are shown in Table 10. Young drivers (<25) have the highest odds ratio, followed by old drivers (> 65). The odds ratio in this model represents the propensity of a driver belonging to a particular age group to be at fault with respect to the reference group. The odds ratios for the age groups follow the typical U-shape curve of crash involvement, with higher probabilities for younger and older drivers. RAIR is the metric used in the quasi-induced exposure and is analogous to the odds ratio. RAIR and probability are measures dependent on the constant in the model, which in turn hinges on estimates for other variables in the model. The probability of all the driver groups being at fault is 1, which means that all groups are equally probable to be at fault. This contradicts the values noted in the odds ratio. The RAIRs depict a U-shape curve as well; however, their values are unrealistic. This could be attributed to the small sample size used in the analysis and the potential lack of variability in the driver age groups data.

Table 10 Ratios of Single-Unit Commercial Vehicle Model

Variable	Estimate	Odds Ratio	Probability, p (at-fault)	RAIR (at-fault)
<25	0	1.000	1.000	2.91E+11
25-39	-0.353	0.703	1.000	2.04E+11
40-64	-0.442	0.643	1.000	1.87E+11
>65	-0.222	0.801	1.000	2.33E+11
Median housing value	-3.34E-05	1.000		
Percent with high school and some college education	-0.463	0.629		
Household median income	6.84E-05	1.000		
Employment by population ratio	0.147	1.158		
Constant	26.396			

The large RAIRs could be due to the small number of crashes utilized, combined with the need to classify them into several categories to account for the effects of socioeconomic variables. To validate this assumption, a separate test is conducted using only the age groups for the single-unit CMVs. This analysis obtains not only the usual U-shape curve for driver age groups observed in priori research, but having more reasonable values and in agreement with prior findings (i.e., values ranging between 0 and 2 are typically observed for the RAIRs) (Table 11). It can be thus concluded that the combination of the small sample size

and socioeconomic variables does not produce a reasonable model, resulting in unrealistically high values for the RAIRs. However, the findings here, based on the statistically sound model, could be used as indicators of potential socioeconomic variables that influence the at-fault status of CMV drivers in crashes and be used as a starting point in future research.

Table 11 Ratios of Single-Unit Commercial Vehicle Model (Age Groups Only)

Variable	Estimate	Wald	p-value	Odds Ratio	Probability, p (at-fault)	RAIR (at-fault)
<25		51.809	0.000	1.000	0.485	0.942
25-39	-0.322	10.610	0.001	0.725	0.405	0.683
40-64	-0.556	34.436	0.000	0.573	0.351	0.540
>65	-0.201	2.420	0.120	0.818	0.435	0.771
Constant	-0.060	0.449	0.503	0.942		

The literature review identified several human factors believed to be a major cause of CMV crashes. Therefore, it is important to test whether the human factors have any relationship with the socioeconomic variables. As noted previously, human factors are categorized into two groups: behavioral and non-behavioral, resulting in a binary dependent variable. Binary logistic regression is used to test them against socioeconomic variables. Using correlation and recursive partitioning analysis, several variables are shortlisted and tested against the binary human factor variable. Percentage rural, age group, median individual income, and employment by population ratio are the most important predictor variables. The FSA tool identified an interaction between median housing value and median individual income which forces the main effect of median housing value into the model. For an interaction term to be included in the model, both variables should be also included alone as main effects. On introducing the interaction variable and the main variables, median individual income becomes insignificant in the model, however model characteristics improved (Table 12). This is just a relative phenomenon and it does not indicate that the variable is not a good predictor. Table 12 shows the best possible model for single-unit truck with the interaction term included. Even though the introduction of the interaction improves the model characteristics, overall performance of the model (Table 13) is not very strong, indicating that socioeconomic factors do not have a strong relationship with the human factors of single-unit CMV drivers.

Table 12 Human Factors Model for Single-Unit Commercial Vehicles

Variable	Estimate	Wald	p-value
>25		2.706	0.439
25-39	-0.049	0.107	0.744
40-64	0.046	0.101	0.75
>65	-0.177	0.787	0.375
Percentage Rural	0.003	4.909	0.027
Median housing value	-1.26E-07	0.000	0.995
Employment by population ratio	-0.015	4.125	0.042
Median individual income	1.89E-05	0.986	0.321
Median housing value × Median individual income	-3.22E-11	0.152	0.696
Constant	-0.156	0.145	0.704

Table 13 Statistics of Human Factors Model for Single-Unit Commercial Vehicle

Criterion	**Values**
Akaike information criterion (AIC)	6155.89
Bayesian information criterion (BIC)	6207.05
Area under Curve (AUC)	0.53
Percentage Correctly predicted	52.4
Chi-Square (DF)	15.77988 (7)
Prob>ChiSq	<0.0001

Two-Unit Commercial Vehicle Model

The process followed for two-unit CMV crashes is the same as the one used for single-unit CMVs. Correlation testing and recursive partitioning analysis are conducted as an initial step towards variable selection. Additional variables belonging to different categories are also tested to finalize the model with the most appropriate set of variables. For two-unit CMVs, several models are tested and the model statistics compared to determine the most appropriate ones. In addition to developing models using only the predictor variables, models with interaction terms are also tested to assess their potential for inclusion. Model comparisons are conducted using the AIC, BIC and AUC to determine their performance. Among all models tested, the one shown in Table 14 performs the best.

Table 14 Model for Two-Unit Commercial Vehicle Crashes

Variable	Estimate	Standard Error	Wald	p-value
>25			68.362	0.000
25-39	-0.425	0.090	22.418	0.000
40-64	-0.060	0.115	0.275	0.600
>65	-0.555	0.086	41.85	0.000
Median housing value	3.17E-06	9.24E-07	11.798	0.001
Household median income	-5.40E-06	4.06E-06	1.768	0.184
Employment by population ratio	0.012	0.004	10.7	0.001
Percent below poverty level	0.018	0.006	8.043	0.005
Household median income × Percent below poverty level	-5.26E-07	1.73E-07	9.258	0.002
Constant	-0.353	0.277	1.626	0.202

This model defines probability of fault as a function of age group, household median income, median housing value, poverty level, and employment by population ratio. A two-way interaction between household median income and employment by population ratio is also included to improve predictive strength. Including the interaction term in the model improves the AUC significantly. All of the variables in the model are significant at the 95 percent confidence level. Model statistics are given in Table 15.

Table 15 Statistics of Two-Unit Commercial Vehicle Models

Criterion	Values
Akaike information criterion (AIC)	12681.6
Bayesian information criterion (BIC)	12745.8
Area under Curve (AUC)	0.561
Percentage Correctly predicted	58
Chi-Square (DF)	110.888 (8)
Prob>ChiSq	<0.0001

The classification model from the recursive partitioning analysis described 'Percent of high school and some college education' seems to be a potential predictor of fault probability. However, including this variable in the model degrades its performance and predictive ability. Models developed with different categories of education, including 'Percent of high school and some college education' drop the AUCs to 0.50. Followed by education, median housing value, employment by population ratio, household median income, and their interaction are other important descriptors in the model. On including the interaction between household median income and percent below poverty level, the main effect of household median income turns insignificant in the model. Perhaps because of its confounding relationship with the interaction term.

The model is similar to that of single-unit CMVs except for the lack of an education and poverty level variable. Recursive partitioning analysis recommends both individual and household income as good predictors. However, the model using the median household income performs better, and it is considered as the best indicator of a driver's economic condition. This is also consistent with the single-unit CMV model and the automobile model (discussed next). Age groups are another predictor variable in the model, and they follow the same pattern as the one observed for the single-unit model. In other words, younger and the older CMV drivers are more likely to cause a crash. The partitioning model results suggest percent never married and household units as the descriptors, however, they are not statistically significant in the logistic model.

The model's intercept is not significant. Generally, the intercept is part of a model and is almost always significantly different from zero. If the intercept is zero (equivalent to having no intercept in the model), the resulting model implies that the response function must be exactly zero when all the predictors are set to zero or at their reference levels. For a logistic model, it means that the logit response function (or log odds) is zero, which implies that an event's probability of occurrence is 0.5. If the intercept is not significant, it should not be removed from the model because it creates a model with the response function equal to zero (i.e., not-at-fault) when the predictors are all zero. Therefore, the constant is included in the model regardless of its larger p-value.

The odds ratio, probability of being the at-fault driver, and RAIR are examined (Table 16). Young drivers (<25) have the highest odds ratio, followed by the middle-aged drivers (ages 40-64). According to prior research, older drivers are more likely to be at fault than the middle-aged drivers. However, this does not hold true here. This could be attributed either to the influence of the model's other socioeconomic variables or the small numbers of drivers in the > 65 category. Analyzing the age groups only as a predictor variable results in the usual U-shaped curve, indicating the potential interaction between sample size and the need to partition the data into several categories for socioeconomic variables. As is the case for the single-unit CMV model, this model can be used to identify potential variables of interest for further research.

Table 16 Ratios for Two-Unit Commercial Vehicle Model

Variable	Estimate	Odds Ratio	Probability, p (at-fault)	RAIR (at-fault)
>25	0.000	1.000	0.413	0.703
25-39	-0.425	0.654	0.315	0.459
40-64	-0.060	0.942	0.398	0.662
>65	-0.555	0.574	0.287	0.403
Median housing value	3.17E-06	1.000		
Household median income	-5.40E-06	1.000		
Employment by population ratio	0.012	1.012		
Percent below poverty level	0.018	1.018		
Household median income × Percent below poverty level	-5.26E-07	1.000		
Constant	-0.353			

The next step is to examine whether the human factors reported for two-unit CMVs crashes have any relationship with the socioeconomic factors. The same two categories are used here as well (i.e., behavioral and non-behavioral human factors). Using correlation and recursive partitioning analysis, several variables are shortlisted and tested against the binary human factor variable. Percentage of high school graduate, age group, and percent separated are the most important predictor variables. The FSA tool identified an interaction between median housing value and median individual income which forces the main effect of median housing value into the model. Introducing the interaction variable and the main effect increases the p-value for percent separated, rendering it insignificant in the model. The model estimates are given in Table 17. Even though introducing the interaction term improves the model characteristics (shown in Table 18), the model's overall performance is not very strong. Again, it can be concluded that socioeconomic factors are not strongly related to the human factors of two-unit CMV drivers.

Table 17 Human Factors Model for Two-Unit Commercial Vehicles

Variable	Estimate	Wald	p-value
>25		1.393	0.707
25-39	0.057	0.233	0.629
40-64	0.166	1.253	0.263
>65	0.049	0.193	0.660
Percentage high school graduate	0.012	5.638	0.018
Median housing value	4.05E-07	0.098	0.755
Percent Separated	-0.038	0.436	0.509
Median housing value × Percent Separated	-3.32E-07	0.467	0.494
Constant	-0.156	0.145	0.704

Table 18 Statistics of Human Factors Model for Two-Unit Commercial Vehicle

Criterion	Values
Akaike information criterion (AIC)	6155.89
Bayesian information criterion (BIC)	6207.05
Area under Curve (AUC)	0.53
Percentage Correctly predicted	52.4
Chi-Square (DF)	15.77988 (7)
Prob>ChiSq	0.0272

5.2 Automobiles

To analyze the relationship between automobile drivers involved in crashes and their socioeconomic characteristics, datasets are prepared following the steps explained in the methodology. Two datasets are generated based on the number of units involved (i.e., single and two) and the crashes are linked to the socioeconomic data of the driver residence. Correlation and recursive partitioning analysis are used to develop a preliminary variable selection. Test results are discussed in the next section. Based on the results from the variable selection process, several models are tested. Model parameters are compared to recommend the most appropriate model for determining the probability of a driver to be at fault.

5.2.1 Variable Selection

Correlation Test

A correlation matrix for single and two-unit automobile crashes is developed to identify variables that are associated with at-fault status. Correlation coefficients are computed for each variable, which represent its association with the dependent variable. Variables statistically significant at 95 percent are shaded green in Table 19.

Correlation analysis for each socioeconomic characteristic identifies those that are significantly related to driver at-fault status. P-values less than 0.05 are regarded as significantly correlated with the at-fault status at the 95 percent level in a two-tailed test. The arithmetic sign of a Pearson coefficient indicates the direction of the relationship between the socioeconomic variable and indicator of crash occurrence. For example, income variables are negatively correlated with at-fault status, meaning that as the income of the driver becomes lower, the likelihood to cause a crash increases.

Among the five categories of race, the proportions of white and black is not significantly correlated with crash occurrence for two-unit crashes, while they are significantly correlated with single-unit crashes. The sign indicates that percent white is negatively correlated with single-unit crashes, which means that drivers from zip codes with more white population are likely to cause fewer single-unit crashes. At the same time, a positive correlation is observed with percent black. In Kentucky, the proportion of people belonging to other races is significantly smaller. Race variables are considered in the statistical modelling to examine whether they are significant when considered alongside other variables.

Table 19 Correlation Analysis Results for Automobile Vehicle Crashes

Category	Variable	Single-Unit		Two-unit	
		Correlation Coefficient	P-value	Correlation Coefficient	P-value
Race	Percent white	-.107**	0.000	0.001	0.674
	Percent black	.092**	0.000	-0.001	0.309
	Percent American Indian	.023**	0.000	.008**	0.000
	Percent Asian	.090**	0.000	-.004**	0.005
	Percent other races	.080**	0.000	.007**	0.000
Housing	Household units	.110**	0.000	-0.001	0.332
	Household ownership total	.112**	0.000	-0.002	0.212
	Owner occupied housing units	.101**	0.000	-.006**	0.000
	Renter occupied housing units	.114**	0.000	.004**	0.009
	Median housing value	.086**	0.000	-.010**	0.000
Marital Status	Percent now married	-.076**	0.000	-.007**	0.000
	Percent widowed	-.052**	0.000	.007**	0.000
	Percent divorced	.012**	0.000	.007**	0.000
	Percent separated	-0.006	0.052	0.003	0.073
	Percent never married	.100**	0.000	.004**	0.005
Education	Percent less than high school graduate	-.086**	0.000	.008**	0.000
	Percent high school graduate	-.093**	0.000	.005**	0.001
	Percent some college or associate degree	.077**	0.000	0.000	0.942
	Percent bachelor's degree or higher	.093**	0.000	-.007**	0.000
	Percent graduate or professional degree	.075**	0.000	-.007**	0.000
Income	Median individual income	.059**	0.000	-.011**	0.000
	Mean individual income	.066**	0.000	-.009**	0.000
	Household mean income	.055**	0.000	-.011**	0.000
	Household median income	.048**	0.000	-.012**	0.000
Other	Employment population ratio	.094**	0.000	-.006**	0.000
	Percentage rural	-.133**	0.000	.003*	0.018
	Unemployment rate	-.025**	0.000	.004**	0.005
	Percent below poverty level	-.030**	0.000	.011**	0.000
	Total population	.109**	0.000	-0.003	0.052

Similarly, housing density has no statistical association with two-unit crashes, while all of the housing variables are related to single-unit crashes. Housing density could be related to rurality and this aspect is tested in the modeling. Housing value is another factor related to rurality. This could be related to household income, as families with high income tend to live in areas with high home values. Housing ownership characteristics (rental/owned) are also correlated with crash occurrence. These relationships are further investigated in the final modeling.

Marital status shows results that agree with prior research. Drivers previously married (widowed, separated, divorced) are correlated with the at-fault status and their crash involvement has been considered as a result

of stressful life events. This is further investigated in the next level of analysis. Furthermore, education shows results in agreement with prior research: less educated people are more likely to be the at-fault driver in a crash. In Table 15, as the educational attainment increases, the sign of the correlation coefficient turns negative which indicates lower crash involvement as an at-fault driver.

All types of income have significant relationships with driver at-fault status according to previous research, but household median income is expected to be a better predictor of crash occurrence (Stamatiadis and Puccini 1999; Lee et al. 2014). These variables indicate a positive relationship with crash occurrence, agreeing with the findings of previous research. The analysis of this research examines various income categories to determine the most appropriate one for inclusion in the final model predicting crash occurrence.

Other variables such as rurality, poverty level, and unemployment rate, which have well-established relationships with crash occurrence, may be also correlated with income and educational level and their interaction is examined.

Logistic regression is the most appropriate and widely used method when the dependent variable is categorical. This method allows more than one independent variable which can be discrete and/or continuous. The dependent variable is the driver fault status, which takes a value of 0 or 1 depending on whether they are at-fault. Based on the correlation test, variables significantly related to crash occurrence are identified. However, this test does not identify the best variables for predicting crash occurrence and combinations of variables should be used to define the most appropriate model. This is a tedious process and the recursive partitioning analysis identifies the best combination of variables that can be used in the model.

Recursive Partitioning Analysis

Based on the regression tree model of two-unit crashes, socioeconomic variables such as age group, gender, percent rural, percent below poverty level, household median income and percent not married now seem to be the most important variables related to crash occurrence. At the same time, percent rural, median household income, unemployment rate, percent white, and percent widowed are the most important variables for describing the occurrence of single-unit crashes. With these results taken into consideration, other variables belonging to different categories (shown in Table 2) are tested to develop different regression models for both single and two-unit crashes. Each variable from the categories are tested to identify the best representation of that category for predicting crash occurrence. As explained in the previous section, multiple variables from the same category are not used in the same model to avoid complimentary effect.

5.2.2 Regression Models

Based on the correlation test and recursive partitioning analysis, socioeconomic variables such as age group, gender, percent rural, percent below poverty level, household median income and percent not married now seem to be the most important variables (in the order presented) for predicting at-fault probability for a crash. Additional variables shown in Table 2 are also tested to determine whether they could be included in the final model. Several models were tested to obtain the most appropriate one for predicting a driver's at-fault probability. The same model parameters (i.e., AIIC, BIC, AUC and Log-likelihood) that were used before are used here as well to select the final model.

Single-Unit Automobile Model

The results from the correlation test and recursive partitioning analysis are used as an initial step toward variable selection for the single-unit model. Similar to the two-unit CMV modeling, several models are tested here. An interaction between median housing value and employment ratio is identified. However, including this term in the model forces the inclusion of the main effects in the model, resulting in lower

model performance. The best performing model for single-unit automobile vehicles is given in Table 20. The large sample size allows for the use of all seven age categories.

Table 20 Model for Single-Unit Automobile Crashes

Variable	Estimate	Wald	p-value
<20		7075.94	0.000
20-24	-0.363	245.613	0.000
25-39	-0.899	1952.04	0.000
40-64	-1.351	4539.19	0.000
65-75	-1.523	2881.7	0.000
75-84	-1.130	847.5	0.000
>84	-0.656	60.181	0.000
Household Median Income	0.000	105.683	0.000
Percent Rural	0.008	1279.47	0.000
Female	-0.511	2153.1	0.000
Percent White	0.007	182.949	0.000
Percent less than high school graduate	0.007	31.05	0.000
Constant	0.374	48.11	0.000

This model defines at-fault probability of a driver as a function of age group, gender, household median income, rurality, education level, and race. All model variables are significant at the 95 percent confidence level.

Age group and gender behave as expected and agree with the findings of prior research. The Wald score is the highest for them, indicating strong associations with at-fault probability for a crash involvement. Percent rural is another important variable with a high Wald score. It is a strong indicator of at-fault probability and agrees with the results of the recursive partitioning analysis.

Percent white and percent widowed are included from the classification model of the recursive partitioning analysis. Percent white turned out to be a useful predictor in the model, however, percent separated is not significant in the logistic model. Attempts to include marital status variables lower overall model parameters.

Several representations of education are tested with the model. 'Percent of less than high school graduate' seems to be the best option among all the variables representing education. The positive coefficient of the education variable indicates that drivers with less than a high school education are more likely to be the at-fault driver in a crash. Education plays a major role in single-unit CMV crashes as well. Single-unit crashes, irrespective of the type of vehicle involved, have a significant relationship with the education level of the drivers in a zip code.

Table 21 shows the characteristics of model. The model has an AUC of 0.68, which indicates that it can adequately distinguish between the two classes (i.e., at-fault and not-at-fault status). The model predicts 66 percent of the validation data correctly, which confirms the quality of the model.

Table 21 Statistics of Single-Unit Automobile Models

Criterion	Values
Akaike information criterion (AIC)	192693
Bayesian information criterion (BIC)	192762
Area under Curve (AUC)	0.68
Percentage Correctly predicted	66
Chi-Square (DF)	14122.55 (6)
Prob>ChiSq	<0.0001

Table 22 shows the odds ratios, probability of being the at-fault driver, and RAIR for the model. Both the odds ratios and RAIR follow the anticipated U-shape curve and agree with prior research findings. Similarly, the driver's gender follows the a priori expectation: male drivers are more likely to be the at-fault driver in a single-unit crash.

Table 22 Ratios of Single-Unit Automobile Model

Variable	Estimate	Odds Ratio	Probability, p (at-fault)	RAIR (at-fault)
<20		1	0.592	1.454
20-24	-0.363	0.696	0.503	1.011
25-39	-0.899	0.407	0.372	0.592
40-64	-1.351	0.259	0.273	0.376
65-75	-1.523	0.218	0.241	0.317
75-84	-1.13	0.323	0.32	0.47
>84	-0.656	0.519	0.43	0.754
Household Median Income	0	1		
Percent Rural	0.008	1.008		
Male		1	0.592	1.454
Female	-0.511	0.6	0.466	0.872
Percent White	0.007	1.007		
Percent less than high school graduate	0.007	1.007		
Constant	0.374			

Two-Unit Automobile Model

Table 23 shows the two significant models for two-unit automobile crashes. These models define the at-fault probability of a driver as a function of age group, gender, household income, poverty level, marital status and employment by population ratio. All of the variables in the model are significant at the 95 percent confidence level. Model 2 includes an interaction term as well.

Table 23 Models for Two-Unit Automobile Crashes

Variable	Model 1			Model 2		
	Estimate	Wald	p-value	Estimate	Wald	p-value
<20		13182.633	0.000		13149.292	0.000
20-24	-0.353	733.655	0.000	-0.353	732.657	0.000
25-39	-0.795	4843.936	0.000	-0.795	4828.942	0.000
40-64	-1.049	8817.574	0.000	-1.049	8799.762	0.000
65-75	-0.776	2858.225	0.000	-0.777	2859.798	0.000
75-84	-0.276	209.164	0.000	-0.277	210.044	0.000
>84	0.174	18.438	0.000	0.171	17.902	0.000
Household median income	-1.10E-06	7.856	0.005	-1.93E-06	11.265	0.001
Percent below poverty level	0.002	11.598	0.001	0.002	5.271	0.022
Female	-0.160	746.325	0.000	-0.160	743.071	0.000
Percent not married now	0.003	7.279	0.007	0.003	7.811	0.005
Employment by population ratio	0.002	15.337	0.000	-2.13E-04	0.073	0.788
Median housing value				-9.09E-07	7.914	0.005
Employment population ratio × median housing value				1.81E-08	10.166	0.001
Constant	0.669	170.090	0.000	0.811	137.138	0.000

Several representations of education are attempted for inclusion in the model. However, their inclusion does not improve the model's predictive strength, indicating that education does not play a major role in determining the probability of the fault status of a driver. The Wald chi-square for age groups and gender are remarkably large, implying their large contribution in predicting the dependent variable. The model represents employment in terms of employment by population ratio, while unemployment rate is also tested. Employment by population ratio is the proportion of the employed population in the total civilian population. At the same time, unemployment rate is the ratio of employed population to the civil labor force. The first displays a better Wald chi-square, improving the model parameters and predictability significantly. Poverty level and marital status also turn out to be important predictors of at-fault status, confirming the results of the recursive partitioning analysis. Similar to the findings of prior research, the model identifies median household income as the best indicator of a driver's economic condition, as it considers the income of all people living in that household. The other representations of individual and household income tested in the model do not improve its predictive ability.

Interactions are also tested using the FSA. A two-way interaction between median housing value and employment by population ratio is identified. Model 2 in Table 23 shows the model incorporating the interaction and the main effect of the variables forming the interaction.

The characteristics of the two models are given in Table 24. The model with the interaction term performs better, however, it is not significant. The AIC and BIC are lower in the new model, implying improved model predictability. However, the AUC, which represents the model's ability to distinguish between the two categories of the predictor variable, remains unchanged. Therefore, to maintain simplicity in the mathematical framework, Model 1 (Table 23) is used.

Table 24 Statistics for Two-Unit Automobile Vehicle Models

Criterion	Model 1	Model 2
Akaike information criterion (AIC)	650188	648689
Bayesian information criterion (BIC)	650321	648845
Area under Curve (AUC)	0.6	0.6
Percentage Correctly predicted	57	57
Chi-Square (DF)	14454.5 (11)	14454.5 (13)
Prob>ChiSq	<0.0001	<0.001

The odds ratio, probability of being the at-fault driver, and RAIR are shown in Table 25. The results are similar to previous research, indicating a U-shaped curve for age groups and higher male involvement as the at-fault driver in two-unit crashes.

Table 25 Ratios for Two-Unit Automobile Vehicle Model

Variable	Estimate	Odds Ratio	Probability, p (at-fault)	RAIR (at-fault)
<20		1.000	0.661	1.953
20-24	-0.353	0.703	0.578	1.372
25-39	-0.795	0.452	0.469	0.882
40-64	-1.049	0.350	0.406	0.684
65-75	-0.776	0.460	0.473	0.898
75-84	-0.276	0.758	0.597	1.481
>84	0.174	1.190	0.699	2.323
Household median income	-1.10E-06	1.000		
Percent below poverty level	0.002	1.002		
Male		1.000	0.661	1.953
Female	-0.160	0.852	0.625	1.663
Percent not married now	0.003	1.003		
Employment by population ratio	0.002	1.002		
Constant	0.669	1.953		

6. Conclusions

This research sought to demonstrate the relationship between crash occurrence and socioeconomic factors associated with the at-fault driver residence, using U.S. Census data. Single and two-unit crashes involving CMVs and automobiles are analyzed separately. Mathematical models are developed that identify socioeconomic characteristics of a driver's home zip code that make them more likely to cause a crash. Key socioeconomic factors considered include income, education level, poverty level, employment, driver age, and the rurality of an area. Several other factors, including marital status and race, are also tested.

In this type of research, it is important to consider crash exposure when analyzing crashes and attempting to identify factors which contribute to crashes. Crash databases do not contain information on driver exposure. The quasi-induced exposure technique is used, which assumes that the not-at-fault drivers represent the total population in question; the crash rate measure of exposure is developed in terms of the relative accident involvement ratio (RAIR). RAIR is the ratio of the percentage of at-fault drivers to the percentage of not-at-fault drivers from the same subgroup. Hence, the dependent variable used here is the fault status of a driver involved in a crash, which is binary.

Logistic regression is used because the dependent variable is categorical. This modeling technique is beneficial when effects of more than one explanatory variable influence an outcome. Independent variables can be discrete and/or continuous, and the response variable is the probability of the outcome; it is modeled based on a combination of the predictor values. Using this technique, regression models for automobile and CMVs are developed as a function of several socioeconomic variables and then compared to learn how they differ within the two vehicle categories.

The model results for the CMV single-unit crashes are quite similar to that of automobile crashes. For CMVs, fault status is a function of age group, household median income, median housing value, education, and employment by population ratio. For automobiles, fault status is linked to age group, gender, household median income, rurality, education, and race. All of the variables in the model are significant at the 95 percent confidence level. Gender is not a significant variable for CMVs, as most of the CMV drivers are males.

The odds ratios for younger and older drivers show greater likelihood of causing a crash for both single-unit CMVs and automobiles. This is consistent with past research, which has shown a relationship between crash involvement and age. Aguero-Valverde et al. (2006) concluded that age groups below 25 and over 65 are positively associated with crash risk, and most of the previous literature has shown a positive association between young drivers and crashes or fatalities. Several studies on older drivers identified their increased crash involvement and demonstrated the risk factors they create for themselves and other drivers. Other studies have also noted that young and old drivers have a positive relationship with crash involvement, indicating their higher propensity to be the at-fault driver in a crash. These are consistent with the findings of this study.

Education plays a major role in single-unit crashes. The CMV model indicates that drivers from a zip code with high school and some college education are less likely to be the at-fault driver, while the automobile model shows that drivers with less than a high school education are more likely to be the at-fault driver. Both models use an education descriptor with the same trends, however, the specific descriptor is different. Previous research has shown a well-defined relationship between level of education and crashes and the findings here are consistent with these. The percentage of people with different education levels and their relationship linked with gender are also significant descriptors of crash propensity (Zephaniah et al. 2018). Hasselberg et al. (2005) determined that drivers with a relatively low levels of educational attainment show an excess risk of overall crashes and of crashes leading to fatality or serious injury. Their study also

estimated that 33 percent of minor injuries and 53 percent of severe injuries would be avoided if all subjects had the same injury rate as subjects with a higher education.

Household median income is another variable common to both models. This agrees with prior findings demonstrating that household income is a better predictor of at-fault status (Stamatiadis and Puccini 1999; Lee et al. 2014). Both income and poverty were cited as relevant predictors for crash-related analysis by other sources, including Aguero-Valverde et al. (2006) who found that the percent of population under the poverty level had a highly significant and positive correlation with crash risk when using a negative binomial prediction model. Median housing value is a good descriptor of single-unit CMV crashes, but none of the housing variables were included in the automobile model, regardless of their significance in the correlation test.

Race is an important factor influencing automobile at-fault involvement; however, there is no significant relationship between race and CMV crashes.

Overall, the models convey similar results and it can be concluded that the socioeconomic factors have similar influence on single-unit CMV and automobile crashes. To confirm these conclusions, two-unit crash models for automobile and CMVs are compared. The CMV model considers at-fault probability of a driver as a function of age group, household median income, median housing value, poverty level, and employment by population ratio. The automobile model includes similar variables: age group, gender, household median income, poverty level, marital status, and employment by population ratio. As expected, gender and age groups show results similar to single-unit crashes. Several categories of education are tested in both models, but education does not have a significant influence on two-unit crashes. Household median income is again the most appropriate descriptor of income in both models, as it was also in the single-unit models. Employment by population ratio and poverty level are additional common variables that show similar effect on both models. The models for CMVs and automobiles look similar. However, marital status (percent not married now) and median housing value have varying influence o CMV and automobile crashes. Both variables correlate with automobile and CMV crashes; however, they are not included in the final models because they were not significant variables in the logistic regression.

Based on the quantitative and qualitative comparison between the models, socioeconomic factors have similar influence on CMV and automobile crashes, irrespective of the number of vehicles involved. The models developed for CMV crashes do not offer the robust statistical support the study was seeking to definitively identify socioeconomic factors that could affect the probability of a CMV driver to be the at-fault one in a crash. This is due to the combination of a small number of crashes where these drivers were involved and the need to classify them in the various socioeconomic variables considered resulting in an even small number for each category. The final result of this was models that were not logical (i.e., unreasonable RAIRs for the drivers). However, once the socioeconomic variables were removed, the RAIRs followed the a priori U-shape curve. These models could still serve as indicators of variables to consider in future research, since they point to their potential contribution, and it was also verified from the automobile models.

References

Abdalla, I. M., Raeside, R., Barker, D. and David R.D, M. (1997). An investigation into the relationships between area social characteristics and road accident casualties. *Accident Analysis and Prevention* 29(5): 583-593.

Adanu, E. K., Penmetsa, P., Jones, S. and Smith, R. (2018). Gendered Analysis of Fatal Crashes among Young Drivers in Alabama. *Safety* 4(3): 29.

Adanu, E. K., 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* 109: 10-17.

A Feasible Solution Algorithm (FSA) for Finding Interactions. shinyfsa.org/. Accessed 5/10/19.

Aguero-Valverde, J. and Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention* 38(3): 618-625.

United Census Bureau. American Census Survey.https://www.census.gov/programs-surveys/acs/. Accessed 3/20/19

Blasco, R. D., Prieto, J. M. and Cornejo, J. M. (2003). Accident probability after accident occurrence. *Safety Science* 41(6): 481-501.

Blatt, J. and Furman, S. M. (1998). Residence Location of Drivers Involved in Fatal Crashes. *Accident Analysis and Prevention* 30(6): 705-711.

Brown, K. T. (2016). A safety analysis of spatial pehonomena about the residences of drivers involved in crashes. *Dissertation Presented to t he Graduate School of Clemson University*.

Cambron, A., Stamatiadis, N., Wright, S. and Sagar, S. (2018). MRI 1: Effect of socioeconomic and demographic factors on kentucky crashes. Southeastern Transportation Center. UT Center for Transportation Research. 309 Conference Center Building. Knoxville TN 37996-4133.

Carr, B. R. A statistical analysis of rural ontario traffic accidents using induced exposure data. *Accident Analysis and Prevention*, Vol. 1, 1969, pp. 33-357.

Chandraratna, S., Stamatiadis, N. and Stromberg, A. (2005). Potential crash involvement of young novice drivers with previous crash and citation records. *Human Performance; Simulation And Visualization*(1937): 1-6.

Chandraratna, S. K. (2004). Crash involvement potential for drivers with multiple crashes. Doctoral, University of Kentucky.

Chandraratna, S., and Stamatiadis, N. Quasi-induced exposure method: Evaluation of not-at-fault assumption. *Accident Analysis and Prevention*, Vol. 2009, No. 41, 2009, p. 2.

Chen, C., Zhang, G., Tian, Z., Bogus, S. M., & Yang, Y. (2015). Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. *Accident Analysis & Prevention*, *85*, 186-198.

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

Chen, H. Y., Ivers, R. Q., Martiniuk, A. L. C., Boufous, S., Senserrick, Woodward, M., Stevenson, M. and Norto, R. (2010). Socioeconomic status and risk of car crash injury, independent of place of residence and driving exposure: results from the DRIVE Study. *Journal of Epidemiology and Community Health* 64(11).

Chen, Q., Williams, S. Z., Liu, Y., Chihuri, S. T., & Li, G. (2018). Multiple imputation of missing marijuana data in the Fatality Analysis Reporting System using a Bayesian multilevel model. *Accident Analysis & Prevention*, *120*, 262-269.

Cook, L. J., Knight, S. and Olson, L. M. (2005). A comparison of aggressive and DUI crashes. *Journal of Safety Research* 36(5): 491-493.

Das, S., Sun, X., Wang, F. and Leboeuf, C. (2015). Estimating likelihood of future crashes for crash-prone drivers. *Journal of Traffic and Tranportation Engineering (English Edition)* 2(3): 145-157.

Factor, R., Mahalel, D. and Yair, G. (2008). Inter-group differences in road-traffic crash involvement. *Accident Analysis and Prevention* 40.

Federal Motor Carrier Safety Administration (2017). Pocket Guide to Large Truck and Bus Statistics, Washington, D.C.

Frost, J. (2019). Understanding Interaction Effects in Statistic, from https://statisticsbyjim.com/regression/interaction-effects/. Accessed 6/10/19.

Gates, J., Dubois, S., Mullen, N., Weaver, B., & Bédard, M. (2013). The influence of stimulants on truck driver crash responsibility in fatal crashes. *Forensic science international*, *228*(1-3), 15-20.

Girotto, E., Mesas, A. E., de Andrade, S. M., & Birolim, M. M. (2014). Psychoactive substance use by truck drivers: a systematic review. *Occup Environ Med*, *71*(1), 71-76.

Hanna, C. L., Laflamme, L. and Bingham, C. R. (2012). Fatal crash involvement of unlicensed young drivers: County level differences according to material deprivation and urbanicity in the United States. *Accident Analysis and Prevention* 45: 291-295.

Harrah, J. (2015). Kentucky Metropolitan Areas Out-Perform Rural and Small Urban Areas. http://crcblog.typepad.com/crcblog/kentucky-metropolitan-areas-out-perform-rural-and-small-urban-areas.html. Accessed 4/10/19.

Hasselberg, M., Vaeza, M. and Laflamme, L. (2005). Socioeconomic aspects of the circumstances and consequences of car crashes among young adults. *Social Science & Medicine* 60(2): 287-295.
Insurance Institute of Highway Safety (2017). Fatality Facts 2017, Highway Loss Data Institute, Washington D.C.

Ivan, J., Burnicki, A., Wang, K. and Mamun, S. (2016). Improvemnts to road safety improvement selection procedures for Connecticut. Connecticut Transportation Institute, Storrs, CT.

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

Knauth, D. R., Pilecco, F. B., Leal, A. F., Seffner, F., & Teixeira, A. M. F. B. (2012). Staying awake: truck drivers' vulnerability in Rio Grande do Sul, Southern Brazil. *Revista de saude publica*, *46*(5), 886-893.

Kocatepe, A., Ulak, M. B., Ozguven, E. E., Horner, M. W. and Arghandeh, R. (2017). Socioeconomic characteristics and crash injury exposure: A case study in Florida using two-step floating catchment area method. *Applied Geography* 87: 207-221.

Kentucky State Police (2016). Kentucky's Traffic Collision Facts, https://transportation.ky.gov/HighwaySafety/Documents/2016_KY_Traffic_Collision_Facts.pdf. Accessed 4/10/19

Kentucky State Police. Kentucky Collision Analysis. http://crashinformationky.org/AdvancedSearch. Accessed 4/10/19.

Kentucky Public Health (2017). Kentucky Racial and Ethnic Distribution Kentucky Public Health. https://chfs.ky.gov/agencies/dph/Documents/2017KYEthnic_Distribution.pdf. Accessed 4/10/19.

Laerd Statistics, Binary Logisitc Regression using SPSS Statistics, https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php. Accessed 5/15/19.

Laerd Statistics, Point-Biserial Correlation using SPSS Statistics, https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php. Accessed 5/15/19.

Lambert, J., Gong, L., Elliott, C.F., Thompson, K., and Stromberg, A. (2018). rFSA: An R Package for Finding Best Subsets and Interactions. *The R Journal* 10:2, 295-308.

Lee, J., Abdel-Aty, M. and Choi, K. (2014). Analysis of residence characteristics of at-fault drivers in traffic crashes. *Safety Science* 68(0): 6-13.

Lemp, J. D., Kockelman, K. M., & Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit models. *Accident Analysis & Prevention*, *43*(1), 370-380.

Lyman, S., Ferguson, S. A., Braver, E. R. and Williams, A. F. (2002). Older driver involvements in police reported crashes and fatal crashes: trends and projections *Injury Prevention* 8(2): 116-120.

Maciag, M. (2014). America's poor neighbohoods plagued by pedestrian deaths. *Governing* Magazine: State and Local Government News for America's Leaders.

Males, M. A. (2009). Poverty as a determinant of young drivers' fatal crash risks. *Safety Research* 40(6): 443-448.

Mir, M. U., Razzak, J. A., & Ahmad, K. (2013). Commercial vehicles and road safety in Pakistan: exploring high-risk attributes among drivers and vehicles. *International journal of injury control and safety promotion*, *20*(4), 331-338.

Muellerman, R. L. and Mueller, K. (1996). Fatal motor vehicle crashes: variations of crash characteristics within rural regions of different population densities. *The Journal of Trauma: Injury, Infection, and Critical Care* 41(2): 315-320.

National Highway Traffic Safety Administration (2013). National Survey of Speeding Attitudes and Behaviors 2011. National Highway Traffic Safety Administration. U.S. Department of Transportation. Washington DC.

National Highway Traffic Safety Administration (2015). The Economic and Societal Impact Of Motor Vehicle Crashes 2010. National Highway Traffic Safety Administration. US Department of Transportation. Washington DC.

Noland, R. B. and Quddus, M. A. (2004). A spatially disaggregate analysis of road casualties in England. *Accident Analysis and Prevention*: 973-984.

National Transportation Safety Board (1990). Fatigue, Alcohol, Other Drugs, and Medical Factors in Fatal-to-the-Driver Heavy Truck Crashes. *Washington DC. NTSB Safety Study*.

PennState Elberly College of Science. Analysis of Discrete Data: Logistic Regression, https://onlinecourses.science.psu.edu/stat504/node/149/. Accessed 5/10/19.

PennState Eberly College of Science. Recursive Partitioning, https://newonlinecourses.science.psu.edu/stat555/node/100/. Accessed 5/10/19.

Rivas-Ruiz, F., Perea-Milla, E. and Jimenez-Puente, A. (2007). Geographic variability of fatal road traffic injuries in Spain during the period 2002–2004: An ecological study. *BMC Public Health* 7(1): 266.

Souza, J. C., Paiva, T., & Reimão, R. (2005). Sleep habits, sleepiness and accidents among truck drivers. *Arquivos de neuro-psiquiatria*, *63*(4), 925-930.

Spainhour, L. K., Brill, D., Sobanjo, J. O., Wekezer, J., & Mtenga, P. V. (2005). Evaluation of traffic crash fatality causes and effects: A study of fatal traffic crashes in Florida from 1998-2000 focusing on heavy truck crashes.

Stamatiadis, N., & Deacon, J. A. (1995). Trends in highway safety: Effects of an aging population on accident propensity. *Accident Analysis & Prevention*, 27(4), 443–459. doi:10.1016/0001-4575(94)00086-2

Stamatiadis, N., and Deacon, J. A. (1997). Quasi-induced exposure: Methodology and insight. *Accident Analysis & Prevention*, 29(1), 37–52.doi:10.1016/s0001-4575(96)00060-7

Stamatiadis, N. and Puccini, G. (1999). Fatal crash rates in the southeastern United States: Why are they higher? *Transportation Research Board*(1665): 118-124.

Sun, X., Das, S. and He, Y. (2014). Analyzing Crash-Prone Drivers in Multiple Crashes for Better Safety Educational and Enforcement Strategies. *Journal of Transportation Technologies* 4(1).

Torre, G. L., Beeck, E. V., Quaranta, G., Mannocci, A. and Ricciardi, W. (2007). Determinants of within-country variation in traffic accident mortality in Italy: A geographical analysis. *International Journal Of Health Geographics* 6(1): 49.

Vachal, K. (2016). Analysis of Risk Factors in Severity of Rural Truck Crashes (No. MPC 16-308), Upper Great Plains Transportation Institute, Fargo, ND.

World Health Organization (2018). Road traffic injuries, http://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed 5/10/19.

Zephaniah, S., Jr., S. J., Smith, R. and Weber, J. (2018). Spatial Dependence among Socioeconomic Attributes in the Analysis of Crashes Attributable to Human Factors. *Analytic Methods in Accident Research (under review)*.

Zwerling, C., Peek-Asa, C., Whitten, P. S., Choi, S.-W., Sprince, N. L. and Jones, M. P. (2005). Fatal motor vehicle crashes in rural and urban areas: Decomposing rates into contributing factors. *Injury Prevention* 11(1): 24-28.