Accident Analysis and Prevention 143 (2020) 105582

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Identifying high-risk commercial vehicle drivers using sociodemographic characteristics

Shraddha Sagar[a,*], Nikiforos Stamatiadis[b], Samantha Wright[c], Aaron Cambron[a]

[a] Department of Civil Engineering, University of Kentucky, 161 Raymond Bldg., Lexington, KY 40506-0281, United States
[b] Department of Civil Engineering, University of Kentucky, 265 Raymond Bldg., Lexington, KY 40506-0281, United States
[c] Department of Civil Engineering, University of Kentucky, 257 Oliver Raymond Bldg., Lexington, KY 40506-0281, United States

ARTICLE INFO

Keywords:
Highway safety
Socioeconomic factors
Commercial motor vehicles
Quasi-induced exposure technique

ABSTRACT

Crash data, from the state of Kentucky, for the 2015-2016 period, show that per capita crash rates and increases in crash-related fatalities were higher than the national average. In an effort to explain why the U.S. Southeast experiences higher crash rates than other regions of the country, previous research has argued the regions unique socioeconomic conditions provide a compelling explanation. Taking this observation as a starting point, this study examines the relationship between highway safety and socioeconomic and demographic characteristics, using an extensive crash dataset from Kentucky. 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.

1. Introduction

According to the World Health Organization (WHO), 1.25 million people die in road traffic crashes each year, while at least 20 million suffer injuries in non-fatal crashes (World health Organization (WHO), 2018). In the U.S., road traffic crashes are a leading cause of death. Kentucky crash data indicate a 10 percent rise in the states fatalities between 2015 and 2016 (from 761 to 834 fatalities); by percentage, this increase was higher than the national average (Green et al. 2017). Kentucky's per capita crash rate also exceeds the national average. In 2016, NHTSA estimated 22.5 crashes per 1,000 people for the U.S. In Kentucky, the rate was 37.3 crashes per 1,000 people.

In 2016, approximately 11.5 million of the 269 million registered vehicles in the U.S. were commercial vehicles (trucks and buses) (Federal Motor Carrier Safety Administration (FMCSA), 2018). Approximately 6.1 million drivers hold a Commercial Driver's License — accounting for 3 percent of licensed drivers. Commercial vehicles (CMVs) were involved in 11.8 percent of fatal crashes in the U.S. in 2016 and 7.4 percent of the non-fatal crashes. Although CMVs typically drive longer distances, their crash involvement is roughly proportional to their exposure. In 2016, the vehicle miles traveled (VMT) by CMVs accounted for 9.1 percent of total VMT in the U.S. (Federal Motor Carrier Safety Administration (FMCSA), 2018). Yet the number of CMV crashes has steadily increased since 2009 (Insurance Institute for Highway Safety (IIHS), 2018). The number of fatalities in CMV crashes is 31 percent higher in 2018 than in 2009 while in 2017 the number of people who died in CMV crashes was 30 percent higher than in 2009 (Insurance Institute for Highway Safety (IIHS), 2017). Most fatalities in CMV crashes are the occupants of passenger vehicles, which is mainly attributed to the vulnerability of people traveling in smaller vehicles.

Several factors are associated with the occurrence of roadway crashes. Addressing the underlying factors which contribute to safety problems will improve overall roadway safety. Previous research has attempted to determine the relationship between drivers residence socioeconomic status and crash rates (Blatt and Furman 1998; Chandraratna, 2004; Chandraratna et al., 2005; Lee et al., 2014). A recent World health Organization (WHO) (2018) study found people from less privileged socioeconomic backgrounds are more likely to be

* Corresponding author.
E-mail addresses: shraddhasagar@uky.edu (S. Sagar), nick.stamatiadis@uky.edu (N. Stamatiadis), sam.wright@uky.edu (S. Wright), aaron.cambron@uky.edu (A. Cambron).

https://doi.org/10.1016/j.aap.2020.105582
Received 28 January 2020; Received in revised form 29 March 2020; Accepted 4 May 2020
Available online 29 May 2020
0001-4575/

involved in crashes, mainly as a result of human errors.

Stamatiadis and Puccini (1999) demonstrated that fatality rates in the Southeast U.S. are consistently higher than fatality rates in other regions. Its distinct socioeconomic characteristics, they contended, help explain the higher rates. Median household income, unemployment rates, educational attainment, and percentage of rural population were selected as useful indicators. The study suggested that socioeconomic data for the zip code in which drivers reside can potentially serve as a surrogate measure when studying high fatality rates.

The total number of traffic collisions has gradually increased in Kentucky. Between 2009 and 2016, CMV crashes increased 27 percent (Green et al., 2017). As the state is located in the Southeast, its socioeconomic characteristics are suspected as a significant contributor to the increasing crash rates. Analyzing the connection between socioeconomic factors and crash occurrence will help to identify the major causes of increasing crash trends, and in turn, identify areas that require more attention to improve roadway safety.

The objective of this study is to examine CMV driver crash involvement and identify high-risk groups using the sociodemographic and socioeconomic attributes of the zip codes where drivers reside (i.e., residence zip codes). The socioeconomic and geodemographic characteristics of residence zip codes are analyzed in tandem with Kentucky historical crash data to identify factors that can serve as indicators of CMV crashes. Findings presented in this paper will help practitioners identify groups of drivers with a high crash-involvement risk factor. It is critical to determine whether the most at-risk drivers have particular demographic characteristics (e.g., age, gender) and where they reside (e.g., zip codes, urban or rural setting). Based on this knowledge, safety programs can be designed to more efficiently target the most at-risk groups (Eguakun et al. 2015).

2. Literature Review

Researchers have explored the influence of socioeconomic and demographic variables on crash rates. Income, poverty, employment, education, rurality, and driver age all appear to have an impact (Stamatiadis and Puccini, 1999; Noland and Quddus, 2004; Factor et al., 2008; Lee et al., 2014; Brown, 2016, Zephaniah et al., 2018). Very little research has been conducted on CMV crashes and socioeconomic effects; most previous studies have focused on the role of alcohol and drugs in CMV crashes (NTSB 1990, Souza et al., 2005; Girotto et al., 2013; Mir et al., 2013).

Rural areas are generally cited as having higher fatality crash rates than urban areas. Muellerman and Mueller (1996) found the CMV crash fatality rate is 44 percent higher in rural areas than in urban areas. Zwerling et al. (2005) found that fatal crash incidence density was more than two times higher in rural areas than in urban areas. Design elements and higher speed limits could be contributing to this risk (Blatt and Furman, 1998). Blatt and Furman also demonstrated that drivers who reside in rural areas or small towns are more likely to be involved in fatal crashes on rural roads. Economic and behavioral factors of the drivers could have an influence on these crash outcomes.

Educational attainment has also been used to explain crash involvement. Research has shown a negative correlation between education level and vehicle crashes. Hasselberg et al. (2005) determined that drivers with relatively low educational attainment are at greater risk of crash involvement (both fatal and serious injury crashes). Zephaniah et al. (2018) demonstrated that driving under the influence (DUI) crash rates (normalized by population) are influenced by employment, income, education, and housing characteristics.

Income, poverty, and employment have been cited as relevant predictors when performing crash-related analysis. Lee et al. (2014) observed a negative correlation between median family income and the number of at-fault drivers. In other words, drivers from lower income communities are more likely to be responsible for a crash. Aguero-Valverde et al. (2006) found a highly significant positive correlation between percentage of the population below the poverty level and crash risk. Factor et al. (2008) demonstrated that non-skilled workers are over-involved in fatal crashes relative to their proportion in the total population of workers. Adanu et al. (2017) found that unemployed drivers have a higher probability of being at fault in a crash (0.23) or a serious injury crash (0.57).

A relationship exists between crash involvement and age as well. Brown et al. (2016) concluded that the 15-19 age group is the most at risk for an injury or fatal crash, followed by the 20-24 age group. Middle-aged drivers (45-54) drivers have the lowest odds of being at fault in a crash. While Factor et al. (2008) found that young or new drivers are more likely to be in a crash and suffer higher fatality rates, there was some variation surrounding the impact of elderly drivers. Lee et al. (2014) determined that a larger proportion of elderly population decreases the likelihood of drivers being at-fault. Aguero-Valverde and Jovanis (2006) concluded that age cohorts below 25 and over 65 are more likely to be involved in a crash. Several studies on older drivers have shown they are involved in more crashes and described the risk factors they create for themselves and others (Aguero-Valverde and Jovanis, 2006; Kocatepe et al., 2017). Studies have also noted that young and old drivers are more likely to be the at-fault driver in a crash.

Eguakun et al. (2015) attempted to define high-risk driver groups based on their geodemographic attributes, driver behavior and traffic collisions. They developed a framework that could help identifying high-risk drivers based on these attributes and that could be targeted in safety messaging programs. Their findings support the notion that age influences traffic collisions with younger drivers having a higher risk to be involved in a crash. Their study notes that agencies can develop customized safety messaging programs to address specific population targets based on known behaviors.

A driver's gender and marital status (separated or widowed) are good predictors of crash occurrence as well. 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 Alabama, Zephaniah et al. (2018) showed that DUI crashes are related to male employment and female educational attainment.

Researchers have typically used logistic regression to address these issues due to the categorical nature of the dependent variable (Factor et al. 2008). Two other useful statistical techniques employed are the point-biserial correlation and recursive partitioning analysis. Point-biserial correlation measures the strength of association between a continuous variable and a binary variable (Laerd Statistics, 2018). Recursive partitioning analysis is a statistical algorithm used for predictive modelling and machine learning and it attempts to correctly classify data along a decision tree by splitting data into subgroups based on the variables at hand (PennState, 2018). This method also helps to identify variables with the greatest predictive power, capture the relative importance of each variable, and specify variables that should be prioritized during logistic regression modeling. Both statistical analyses help reduce the number of explanatory variables that need to be considered in a model. Interactions among variables often need to be examined in model development. A tool called *Shiny* that uses a Feasible Solution Algorithm (FSA) for identifying appropriate interactions can be used in identifying such interactions (Lambert et al., 2018). FSA allows to identify any order interactions, and this allows users to formulate new models or improve existing models by including the best possible interaction(s).

In sum, previous research has indicated the socioeconomic factors most implicated in crash risk are income, education level, poverty percentage, employment, driver age, and where a driver is from (e.g., rural or urban setting). Typically, education and income are negatively correlated with crash risk, poverty is positively correlated, and the effects of employment vary across studies. Young drivers, and areas with a high proportion of young drivers, tend to have higher crash rates and more fatalities, while the severity of crashes in more rural areas is more pronounced. There has been very limited research on the potential

influence of these socioeconomic and demographic aspects on CMV crashes, making this research a first step towards examining their possible association.

3. Crash and Socioeconomic Data

Four years (2013-2016) of CMV crash data, aggregated by zip code, were supplied by the Kentucky State Police (KSP). During the four-year period, 13.7 percent were single-unit crashes, 77.0 percent of crashes were two-unit crashes, and the rest involved three or more vehicles. Analysis centered on data for 2,955 CMV and 74,751 automobile drivers in single-unit crashes and 9,248 (4,450 in the at-fault group and 4,798 in the not-at-fault group) CMV and 479,444 (239,722 crash pairs) automobile drivers in two-unit crashes. Focusing on single- and two-unit crashes limited the number of drivers involved to a maximum of two. Information on crashes, vehicles, and drivers were extracted from the KSP database.

Human factors are a term used to denote the action of a driver prior to the crash that could identify their culpability in the crash occurrence. The police officer investigating the crash records them for every driver engaged in the crash and they are used to determine at-fault status. Drivers with a human factor code recorded by the police officer were treated as being at fault (Chandraratna and Stamatiadis, 2009), whereas drivers with a factor coded as non-detected were treated as not being at-fault. Researchers eliminated crashes from the analysis if they recorded a human factor code for both drivers and neither driver. This selection criterion avoided designating multiple at-fault drivers for the same crash in two-unit crashes (Chandraratna and Stamatiadis, 2009). For single-unit crashes, only drivers with a human factor coded were included in analysis; these drivers are coded as at-fault. Single-unit crashes do not include a not-at-fault driver group. The not-at-fault driver group from the two-unit crashes were used as the exposure metric (Stamatiadis and Deacon, 1997).

The final datasets included drivers between the ages of 15 and 90. Drivers were grouped into seven age categories: 20 < , 20-24, 25-39, 40-64, 65-74, 75-84 and > 85. For CMVs, only a small number of drivers belonged in the youngest and the oldest groups. As such, these groups were combined to facilitate more statistically meaningful analysis

Socioeconomic and demographic data were collected from the American Census Bureau (ACS) (U.S. Census Bureau, 2016). Table 1 shows the list of Information retrieved from the ACS database. The data for the demographic and socioeconomic descriptors is joined at the zip code level which is then merged to the data of crash related variables matching the residence zip code of the driver. The variables in the final dataset are tested with the dependent variable (at-fault status of the driver) to understand their potential correlation with each other.

4. Methodology

The main objective of this study is to examine CMV driver crash involvement and identify high-risk groups using the sociodemographic and socioeconomic characteristics of the zip codes in which at-fault drivers reside. To analyze the factors contributing to crash occurrence, crash-related factors, driver characteristics, and socioeconomic and demographic features of the residence zip code were all examined. Logistic regression is used here for modelling due to the categorical nature of the dependent variable. As a first step in variable selection, point-biserial correlation and recursive partitioning analysis were conducted. Both statistical analyses help reduce the number of explanatory variables that need to be considered in a model; their results were used as a starting point for the development of logistic regression models.

The point-biserial correlation allows for the identification of any correlation between the dependent variable and the independent predictors. This allows for a preliminary identification of variables that could be considered as predictors in the final model. The recursive partitioning attempts to correctly classify the data along a decision tree, by splitting it in subgroups based on the variables at hand. This method examines all the variables in the dataset to find those that give the best homogeneous group when splitting the data. The tree classification provides a relative understanding of the importance of a variable for its inclusion in the model and this information is used for final variable selection and logistic regression modeling of the drivers at-fault probability. Recursive partitioning is useful for shifting through a large set of variables to identify those that are most likely to be good predictors and this advantage was the main reason for utilizing this approach.

Along with variables identified by these analyses, other variables were tested to finalize models containing the most appropriate set of explanatory variables. Multiple variables from the same category were not used in the same model in order to avoid collinearity. The research team developed several models for single and two-unit crashes. Final models were parameterized based on the contributions and significance of explanatory variables.

Interactions among socioeconomic variables may influence crash occurrence. Some of the potential interaction terms were identified based on prior knowledge and screened individually. FSA was employed to identify two-way interactions to be included in this study. Several criterion functions (e.g., R^2 and adjusted R^2, interaction p-values, Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)) were used to appraise model quality.

4.1. Crash Exposure – Quasi-Induced Exposure Technique

When studying crashes and the factors which contribute to their

Table 1
List of Socioeconomic variables.

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

Table 2
Point-Biserial Correlation Test Results for CMV and Automobile Crashes.

Category	Variable	CMV				Automobile			
		Single-Unit		Two-unit		Single-Unit		Two-unit	
		Corr. Coeff.	P-value	Corr. Coeff.	P-value	Corr. Coeff.	P-value	Corr. Coeff.	P-value
Race	Percent white (WH)	−0.007	0.563	−0.035	0.001	−0.107	0.000	0.001	0.674
	Percent black (BL)	0.008	0.461	0.031	0.003	0.092	0.000	−0.001	0.309
	Percent American Indian (AI)	0.009	0.413	−0.002	0.817	0.023	0.000	0.008	0.000
	Percent Asian (AS)	−0.002	0.860	0.049	0.000	0.090	0.000	−0.004	0.005
	Percent other races (OR)	−0.006	0.608	0.009	0.383	0.080	0.000	0.007	0.000
Housing	Household units (HH)	−0.014	0.608	0.034	0.001	0.110	0.000	−0.001	0.332
	Household ownership total (HHO)	−0.018	0.122	0.034	0.001	0.112	0.000	−0.002	0.212
	Owner occupied housing units (OHU)	−0.032	0.005	0.030	0.004	0.101	0.000	−0.006	0.000
	Renter occupied housing units (RHU)	0.004	0.722	0.035	0.001	0.114	0.000	0.004	0.009
	Median housing value (HVL)	−0.041	0.000	0.043	0.000	0.086	0.000	−0.010	0.000
Marital Status	Percent now married (MRD)	−0.018	0.104	−0.022	0.034	−0.076	0.000	−0.007	0.000
	Percent widowed (WID)	0.030	0.008	−0.020	0.059	−0.052	0.000	0.007	0.000
	Percent divorced (DIV)	−0.002	0.869	−0.003	0.773	0.012	0.000	0.007	0.000
	Percent separated (SEP)	0.012	0.282	−0.004	0.709	−0.006	0.052	0.003	0.073
	Percent never married (NMD)	0.011	0.319	0.032	0.002	0.100	0.000	0.004	0.005
Education	Percent less than high school graduate (LHS)	0.038	0.001	−0.031	0.003	−0.086	0.000	0.008	0.000
	Percent high school graduate (HS)	−0.020	0.086	−0.046	0.000	−0.093	0.000	0.005	0.001
	Percent some college/associate degree (COL)	−0.011	0.334	0.019	0.071	0.077	0.000	0.000	0.942
	Percent bachelors degree or higher (BS)	−0.011	0.315	0.045	0.000	0.093	0.000	−0.007	0.000
	Percent graduate or professional degree (GD)	0.004	0.692	0.041	0.000	0.075	0.000	−0.007	0.000
Income	Median individual income (MDIINC)	−0.050	0.000	0.033	0.001	0.059	0.000	−0.011	0.000
	Mean individual income (MIINC)	−0.034	0.003	0.031	0.003	0.066	0.000	−0.009	0.000
	Household mean income (MHINC)	0.000	0.439	0.026	0.012	0.055	0.000	−0.011	0.000
	Household median income (MDHINC)	−0.049	0.000	0.022	0.037	0.048	0.000	−0.012	0.000
Other	Employment population ratio (EMP)	−0.061	0.000	0.039	0.000	0.094	0.000	−0.006	0.000
	Percentage rural (RUR)	0.019	0.088	−0.045	0.000	−0.133	0.000	0.003	0.018
	Unemployment rate (UEMP)	0.027	0.016	−0.003	0.754	−0.025	0.000	0.004	0.005
	Percent below poverty level (POV)	0.054	0.000	−0.010	0.336	−0.030	0.000	0.011	0.000
	Total population (POP)	−0.021	0.060	0.032	0.002	0.109	0.000	−0.003	0.052

occurrence, it is critical to consider crash exposure. Typically, exposure has been defined using exogenous factors, such as VMT, number of licensed drivers, and registered vehicles. With these conventional exposure metrics, the proportion of the driving population classified as exposed varies depending on other factors, such as time of day, driver gender or age, and road type. This has raised questions about their reliability for examining safety issues. The problem is that they pertain to more specific groups of drivers or conditions, because the denominator in the ratio of crash occurrence for such subgroups and conditions cannot be obtained. Carr (1969) developed the quasi-induced exposure technique to overcome this problem. The technique assumes that the distribution of not-at-fault drivers reflects the distribution of all drivers exposed to the risk of crash involvement. The crash-rate measure of exposure is developed in terms of the relative accident involvement ratio (RAIR). This metric measures the relative crash propensity of a driver group as a ratio of the proportion of at-fault drives to not-at-fault drivers. This risk ratio is analogous to the odds ratio which is also a measure of association between an exposure and an outcome. Stamatiadis and Deacon (1997) demonstrated the method is appropriate for estimating exposure, while Chandraratna and Stamatiadis (2009) validated the assumptions of the approach. They concluded that "estimating relative crash propensities for any given driver type by using the quasi-induced exposure approach will yield reasonable estimates of exposure."

4.2. Statistical Modeling

Logistic regression is the most appropriate and widely used method for modeling categorical response variables. It is beneficial when the effects of more than one explanatory variable influence an outcome (Das et al., 2015). Explanatory variables can be discrete or continuous. In logistic regression, the expected values of the response variable are modeled based on the probability or odds of the response taking a particular value utilizing a combination of predictor values.

The logarithm of odds (i.e., log-odds or logit function) is defined as:

$$\text{logit (p)} = \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{\text{probability of being at}-\text{ fault}}{\text{probability of not being at}-\text{ fault}}\right) \quad (2)$$

In Equation 2 the ratio of the probability of being at-fault to the probability of not being at-fault is equivalent to the relative accident involvement ratio (RAIR), which is the crash rate measure in the quasi-induced exposure technique. The probability is computed as:

$$p = \frac{1}{1 + e^{-f(X)}} \quad (3)$$

where $f(X) = \text{a} + b_1X_1 + b_2X_2 + \ldots + b_nX_n$ is the regression model, X_i is the i^{th} explanatory variable, a is the intercept, and b_i is the i^{th} coefficient estimated using the maximum likelihood method. Target groups (i.e., age-group and gender) and target areas (i.e., zip codes) with high crash propensity can be identified based on the probabilities and RAIRs (or odds ratio) developed using logistic regression.

AIC and BIC were used for the initial model comparisons. They are used to estimate the relative quality of statistical models for a given set of data 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 of an increase in likelihood with the addition of more parameters, which may result in overfitting. Evaluating the receiver operating characteristic (ROC) curve was another method used for model assessments. It is a plot that illustrates the performance measurements of a model. The area under the curve (AUC) represents the degree or measure of separability between two classes.

Training and validation were conducted to calculate the percentage of observations correctly predicted by each model. In general, a training set is larger than a validation set, as this ensures that the training set is a good representation of the overall dataset. In this study, 80 percent of the data were placed in the training dataset and 20 percent were used in

the validation dataset.

5. Statistical Modelling

The point biserial correlation test conducted identified variables associated with at-fault status. Table 2 lists in bold the statistically significant correlations ($p < .05$).

Among the five categories of race, the proportion of white, black and Asians seems to be significantly correlated with two-unit truck-related crashes. Also, no predominant relationship is observed between races and single-unit truck crash occurrence. A different trend is observed for automobile crashes. The proportion of white and black seems to not be significantly correlated with crash occurrence for two-unit crashes, while they are significantly correlated to single-unit crashes. Race does not seem to be an important descriptor of at-fault status of a driver involved in two-unit crashes, although there is a significant association between race and single-unit crashes.

All housing variables are statistically associated with two-unit truck related crashes; however, housing density does not seem related to single-unit truck crashes. For the automobile crashes, housing density failed to show a statistical association to two-unit crashes. At the same time, housing value is correlated to both automobile and truck related crashes regardless of the number of units involved. Housing value could be related to rurality and to household income, as families with high income tend to live in areas with high housing value. Housing ownership characteristics (rental/owned) also seem to be correlated with single-unit and two-unit automobile and truck related crashes, while rented house density is not related to the occurrence of single-unit truck crashes.

Marital status does not seem to have substantial effects on two-unit and single-unit truck crashes, while they seem to have a significant association with automobile crash occurrence. Percent now married and percent never married are significant 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 conducted in the next level of analysis. Furthermore, education seems to be a potential descriptor of two-unit truck related crashes, while it does not seem to have a significant effect on single-unit truck crashes. Education levels showed results in agreement with prior research in case of two-unit automobile crashes: driver sin areas with lower education attainment are more likely to be the at-fault driver in a crash. Percent less than high school graduate seems to be a strong indicator of education which is correlated to the occurrence of all the crash types. Again, the relationship of education with occurrence of automobile and truck related crashes requires more investigation.

Individual, as well as household, income show significant relationship with the at-fault status of the truck as well as automobile driver. Prior research (Stamatiadis and Puccini, 1999; Lee et al., 2014) demonstrated household income to be a better predictor of crash occurrence. The various income categories will be further examined to determine the most appropriate one for inclusion in the final models for predicting occurrence of automobile and truck related crashes. Also, other variables such as rurality, poverty level, and employment by population ratio, that have well established relationship with both automobile and truck related crashes, may be also correlated to income and educational level and their interaction will be examined in the next step.

5.1. Single-unit CMV model

Model 1, which is focused on single-unit CMV crashes, is the simplest one for estimating at-fault driver propensity based on socioeconomic factors (Table 3). The model defines probability of fault as a function of age-group, household median income (MDHINC), median housing value (HVL), an education indicator (percent of high school and some college education, HS + COL), and employment by population ratio (EMP). Model 2 includes an interaction term as well.

The research team examined several education indicators to determine which generated the strongest model; percent of high school and some college education showed the best performance. Models with other education indicators lowered the AUCs to 0.50. The Wald chi-square for this education category was large and its inclusion improved the model's predictive power. Employment by population ratio (EMP), median housing value (HVL), and household median income (MDHINC) are the other socioeconomic descriptors in the model. Recursive partitioning analysis identified both individual and household income as good explanatory variables. However, the model using median household income seemed to be the best indicator of a driver's economic status, which is also consistent with previous research findings. Younger and older drivers are more likely to be the at-fault driver in a crash. Percent white and percent widowed were also recommended from the classification model, however, because they had high p-values and their inclusion lowered AUCs and AIC/BIC, they were omitted. A two-way interaction between income and employment by population ratio was also identified and incorporated into Model 2 (Table 3).

Model parameters were compared, and the model's likelihood functions improved with the inclusion of the interaction term. The improvement was not significant, however. The AUC remained more or less the same. Therefore, to maintain simplicity, Model 1 (AIC = 4,721.6; AUC = 0.943; percent correctly predicted = 88.6) is preferred.

Young drivers (< 25) have the highest odds ratio (or relative risk to be at-fault), followed by the old drivers (> 65). The odds ratios for age groups follow the typical U-shape curve of crash involvement, with younger and older driver having higher probabilities of being in crashes. RAIR is used in the quasi-induced exposure technique and is analogous to the odds ratio, but it uses Equation 3 for estimating the at-fault probability. It shows the probability of all driver groups to be at-fault is 1, which means that all groups are equally likely to be at-fault, contradicting the odds ratios. The RAIRs follow a U-shape curve as well, but their values are not plausible. A separate test was conducted using only the age groups. This analysis returned more reasonable values that agreed with prior research. The combination of the small sample size and socioeconomic variables did not yield a reasonable model, resulting in unrealistically high values for the RAIRs.

5.2. Single-unit Automobile Model

Table 4 shows the best performing model for single-unit automobile crashes. This model (AIC = 192,693; AUC = 0.68; percent correctly predicted = 66) defines at-fault probability of a driver as a function of age, gender, household median income (MDHINC), percent rural (RUR), education level (percent less than high school graduate, LHS), and race (percent white, WH). All explanatory variables are significant ($p < .05$). An interaction between median housing value and employment ratio was identified. However, when an interaction term is included in a model, the main effects must be also included. The large sample size allowed for the use of all seven age groups.

The effects of age group and gender are consistent with previous research findings. The Wald score indicates they are strongly associated with at-fault probability. Percent rural is another variable with a high Wald score. Percent white and percent widowed were included from the classification model of the recursive partitioning analysis. While percent white turned out to be a useful predictor, percent widowed did not. Attempts to include marital status indicators resulted in lower overall model parameters. Among education indicators, percent of less than high school graduate offered the best performance. The positive coefficient of education variables 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.

The models AUC of 0.68 indicates a good ability to distinguish the

Table 3
Models for Single unit CMV crashes.

Variables	Model 1				Model 2			
	Estimate	Wald	p-value	Odds ratio	Estimate	Wald	p-value	Odds ratio
< 25		10	0.019	1		9.96	0.019	1
25-39	−0.353	5.03	0.025	0.703	−0.331	4.18	0.041	0.718
40-64	−0.442	8.59	0.003	0.643	−0.457	8.67	0.003	0.633
> 65	−0.222	1.21	0.272	0.801	−0.28	1.81	0.178	0.756
HVL	−3.34E-05	312.55	0.000	1	−3.39E-05	269.17	0.000	1
HS + COL	−0.463	1777	0.000	0.629	−0.483	1786.01	0.000	0.617
MDHINC	6.84E-05	127.38	0.000	1	3.11E-04	226.23	0.000	1
EMP	0.147	429.11	0.000	1.158	0.313	439.89	0.000	1.368
EMP × MDHINC					−4.21E-06	150.897	0.000	1
Constant	26.396	1653.49	0.000		18.549	452.782	0.000	
Model Parameters								
AIC	4,721.6				4,484.52			
AUC	0.943				0.948			
Percentage Correctly predicted	88.6				88.9			

Table 4
Model for Single unit Automobile Crashes.

Model 3				
Variables	Estimate	Wald	p-value	Odds ratio
< 20		7075.94	0.000	1
20-24	−0.363	245.61	0.000	0.696
25-39	−0.899	1952.04	0.000	0.407
40-64	−1.351	4539.19	0.000	0.259
65-75	−1.523	2881.7	0.000	0.218
75-84	−1.13	847.5	0.000	0.323
> 84	−0.656	60.18	0.000	0.519
MDHINC	−5.31E-06	105.68	0.000	1
RUR	0.008	1279.47	0.000	1.008
Female	−0.511	2153.1	0.000	0.6
WH	0.007	182.95	0.000	1.007
LHS	0.007	31.05	0.000	1.007
Constant	0.374	48.11	0.000	
Model Parameters				
AIC	192,693			
AUC	0.68			
Percentage Correctly predicted	66			

Table 5
Model for Two-unit CMV crashes.

Model 4				
Variables	Estimate	Wald	p-value	Odds ratio
25 <		68.362	0.000	1
25-39	−0.425	22.418	0.000	0.654
40-64	−0.06	0.275	0.600	0.942
> 65	−0.555	41.85	0.000	0.574
HVL	3.17E-06	11.798	0.001	1
MDHINC	−5.40E-06	1.768	0.184	1
EMP	0.012	10.7	0.001	1.012
POV	0.018	8.043	0.005	1.018
MDHINC × POV	−5.26E-07	9.258	0.002	1
Constant	−0.353	1.626	0.202	
Model Parameters				
AIC	12,681.6			
AUC	0.561			
Percentage Correctly predicted	58			

at-fault status of a driver, and it offers correct predictions 66 percent of the time using validation data. The odds ratio and RAIR followed the anticipated U-shape curve and agree with prior research findings. Similarly, the drivers gender follows the a priori expectation: male drivers are more likely to be at fault in a single-unit crash.

5.3. Two-unit CMV Model

The model (AIC = 12,681.6; AUC = 0.561; percent correctly predicted = 58) for two-unit CMV crashes (Table 5) defines probability of fault as a function of age, household median income (MDHINC), median housing value (HVL), poverty level (POV) and employment by population ratio (EMP). The model includes a two-way interaction between household median income and employment by population ratio (MDHINC × POV), which improved its predictive quality. All variables included in the model are significant ($p < .05$).

While recursive partitioning analysis indicated it would be appropriate to include an indicator of education, efforts to add one did not improve the model performance and none were eventually retained. Median housing value, employment by population ratio, household median income, and the interaction of the last two variables are important explanatory variables in the model. Adding a two-way interaction between household median income and percent below poverty level rendered the effect of household median income insignificant. One explanation for this is that household median income is confounded by the interaction term.

Recursive partitioning analysis also recommended both individual and household income as good explanatory variables. However, consistent with the single-unit CMV models, the model using median household income performed better and was considered as the best indicator of a driver's economic status. Age groups are also significant predictors and they follow the same pattern as that of the single-unit CMV models. The partitioning model suggested as candidate variables percent never married and household units, but they were not statistically significant when added to the logistic regression model.

Researchers also examined odds ratios, the probability of being the at-fault driver, and RAIR. Younger drivers (< 25) have the highest odds ratio, followed by the middle-aged drivers (ages 40-64). Previous research has found that older drivers are more likely than middle-aged drivers to be the driver at-fault. 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 number of drivers in the > 65 category. Analysis of the age groups only as a predictor variable resulted in the usual U-shape curve, indicating a potential interaction between sample size and the need to partition the data for socioeconomic variables into several categories.

5.4. Two-unit Automobile Model

Table 6 presents the two significant models for two-unit automobile crashes. These models define the at-fault probability of a driver as a function of age, gender, household income (MDHINC), poverty level (POV), marital status (percent not married now, NMN), and employment by population ratio (EMP). All variables are significant ($p < .05$).

Table 6
Model for Two-unit Automobile crashes.

Variables	Model 5				Model 6			
	Estimate	Wald	p-value	Odds ratio	Estimate	Wald	p-value	Odds ratio
< 20		13182.6	0.000	1		13149.3	0.000	1
20-24	−0.353	733.66	0.000	0.703	−0.353	732.66	0.000	0.703
25-39	−0.795	4843.94	0.000	0.452	−0.795	4828.94	0.000	0.452
40-64	−1.049	8817.57	0.000	0.35	−1.049	8799.76	0.000	0.350
65-75	−0.776	2858.23	0.000	0.46	−0.777	2859.8	0.000	0.460
75-84	−0.276	209.16	0.000	0.759	−0.277	210.04	0.000	0.758
> 84	0.174	18.44	0.000	1.19	0.171	17.9	0.000	1.186
MDHINC	−1.10E-06	7.86	0.005	1	−1.93E-06	11.27	0.001	1
EMP	0.002	15.34	0.000	1	−2.13E-04	0.07	0.788	1
POV	0.002	11.6	0.001	1	0.002	5.27	0.022	1
Female	−0.16	746.33	0.000	0.852	−0.16	743.07	0.000	0.852
NMN**	0.003	7.28	0.007	1	0.003	7.81	0.005	1.000
HVL					−9.09E-07	7.91	0.005	1
EMP × HVL					1.81E-08	10.17	0.001	1
Constant	0.669	170.09	0.000		0.811	137.14	0.000	
Model Parameters								
AIC	650,188				648,689			
AUC	0.6				0.6			
Percentage Correctly predicted	57				57			

** NMN = WID + DIV + SEP

Model 6 includes an interaction term as well.

Several education indicators were tested for potential inclusion in the models. However, none improved their predictive power, indicating that education does not significantly influence probability of being at fault. The Wald scores for age groups and gender are remarkably high, confirming their significance for predicting the response variable. Employment by population ratio and unemployment rate were tested, with the former having a higher Wald score; it was thus included in the model. Poverty level and marital status are important predictors of at-fault status, confirming the results of the recursive partitioning analysis. Median household income is the best indicator of a driver's economic status, aligning with prior research findings. Tests of individual and household income showed they did not improve the model's predictive ability. A significant two-way interaction between median housing value and employment by population ratio was identified. Model 6 incorporates this interaction term as well as the main effect of the variables constituting the interaction.

A comparison indicates that Model 6 performs better than Model 5. AIC and BIC values for Model 6 are lower than Model 5, denoting better predictive power and less information loss. However, the AUCs are unchanged. Thus, to maintain simplicity, Model 5 (AIC = 650,188; AUC = 0.60; percent correctly predicted = 57) is adopted. The odds ratio and RAIR show results similar to past research, returning a U-shape curve for age groups and higher male involvement as the at-fault driver in two-unit crashes.

6. Discussion and Conclusions

This research studied the relationship between crash rates and the socioeconomic attributes of zip codes in which at-fault drivers reside. Single- and two-unit crashes involving CMVs and automobiles were analyzed separately. Models were developed to identify which socio-economic characteristics of a driver's home zip code make them more likely to cause a crash. The quasi-induced exposure technique was used, which assumes not-at-fault drivers represent the total population in question; the crash rate measure of exposure was developed in terms of the RAIR. The dependent variable was the at-fault status of a driver involved in a crash, which is binary. The models were then compared to learn how they differ within the two vehicle categories. The main socioeconomic factors identified as significant included income, education level, poverty level, employment, drivers age, and rurality. Several other factors, including marital status and race, were also tested.

Model results for single-unit CMV crashes were quite similar to those for automobile crashes. For CMV crashes, fault status is a function of age, household median income, median housing value, education and employment by population ratio. For automobiles, fault status is linked to age, gender, household median income, rurality, education, and race. Gender is not a significant variable for CMV crashes, as most of the CMV drivers are males. Odds ratios show that younger and older drivers are more likely to be the at-fault driver in single-unit crashes, for CMVs and automobiles alike. This finding matches up with previous research, which has demonstrated a relationship between crash involvement and age, specifically that younger and older drivers have a greater propensity to be the at-fault driver in a crash (Aguero-Valverde and Jovanis, 2006).

Educational attainment seemingly has a major influence on single-unit crashes. For CMV crashes, high school graduates with some college experience are less likely to be the at-fault driver in a crash. Similarly, drivers who have not graduated from high school are more likely to be the at-fault driver in an automobile crash. Both the CMV and automobile models integrate an education indicator with the same trends, however, the indicator used differs. The findings of the study concur with previous research that has established a well-defined relationship between educational attainment and the likelihood of being the at-fault driver in a crash.

Another variable common to both models is household income. Consistent with earlier research, this study confirms that household income is a better predictor of at-fault involvement in a crash than any other income indicators (Stamatiadis and Puccini, 1999; Lee et al., 2014). Median housing value seems to help explain at-fault status for single-unit CMV crashes, but no housing indicators are included in the automobile model, regardless of their significance in the correlation test.

Two-unit crash models for automobiles and CMVs were compared. Explanatory variables in the CMV model include age, household median income, median housing value, poverty level, and employment by population ratio. The automobile model incorporates similar variables: age, gender, household median income, poverty level, marital status, and employment by population ratio. As expected, the effects of gender and age groups are similar to those observed for single-unit crashes. Several education indicators were tested in both models, but education appears to not significantly influence two-unit crashes. Among possible income-related variables, household median income is the most appropriate indicator of income in both models. Employment

by population ratio and poverty level have similar effects on both models. Overall, the models for CMVs and automobiles look similar. However, marital status (percent not currently married) and median housing value have varying influence on CMV and automobile crashes. Both variables are correlated with automobile and CMV crashes; however, they were omitted from the final models because they were not significant in the logistic regression.

Qualitative and quantitative comparisons of the models conclusively demonstrate that the influence of socioeconomic factors on CMV and automobile crashes is similar, irrespective of how many vehicles are involved in a crash. Predictive power of the CMV models is low due to unexplained variation in the data. However, these models can still help researchers decide what variables to examine in future studies, as they point toward their potential contributions, and as their significance was verified for the automobile models.

The findings of this study will help practitioners identify CMV groups of drivers with a high crash-involvement risk factor. Based on this knowledge, safety programs can be designed to more efficiently target the most at-risk groups. Virtual driving simulators can be a cost-effective way to train the at-risk drivers about the possible real-world dangerous situations they may need to tackle. The 2009 Driver Training Study of California Commission (2009) showed that driver training utilizing a driving simulator results in nearly a 10 percent reduction of traffic collisions. Also, traffic safety messages can be delivered to the target groups to improve awareness (Eguakun et al., 2015). The target groups can be given compulsory safety awareness classes for driver's license renewal. As a control measure to prevent crashes, drivers in at-risk groups can be issued severe penalties (such as license suspension or revocation) if found guilty of a traffic violation or being at fault in a crash.

CRediT authorship contribution statement

Shraddha Sagar: Data curation, Software, Visualization, Formal analysis, Writing - original draft. **Nikiforos Stamatiadis:** Conceptualization, Methodology, Supervision, Formal analysis, Writing - review & editing. **Samantha Wright:** Conceptualization, Methodology, Writing - review & editing. **Aaron Cambron:** Data curation, Writing - original draft.

Acknowledgement

The authors would like to acknowledge the support of the Kentucky Injury Prevention Center, the Kentucky Transportation Cabinet, and the Centers for Disease Control and Prevention. This journal article was supported by Cooperative Agreement Number 5 U60OH008483-15-00, funded by CDC. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or the Department of Health and Human Services.

References

Adanu, E.K., Smith, R., Powell, L., Jones, S., 2017. Multilevel analysis of the role of human factors in regional disparities in crash outcomes. Accident Analysis and Prevention 109, 10–17.
Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38 (3), 618–625.
Blatt, J., 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 phenomena about the residences of drivers involved in crashes. Dissertation Presented to the Graduate School of Clemson University.
California Commission, 2009. Post driver training study. California Commission on Peace Officer Standards and Training.Post driver training study. California Commission on Peace Officer Standards and Training.
Carr, B.R., 1969. A statistical analysis of rural Ontario traffic accidents using induced exposure data. Accident Analysis and Prevention 1, 33–357.
Chandraratna, S.K., 2004. Crash involvement potential for drivers with multiple crashes. University of Kentucky.
Chandraratna, S., Stamatiadis, N., 2009. Quasi-induced exposure method: Evaluation of not-at-fault assumption. Accident Analysis and Prevention 2009 (41), 308–313.
Chandraratna, S., Stamatiadis, N., Stromberg, A., 2005. Potential crash involvement of young novice drivers with previous crash and citation records. Human Performance; Simulation And Visualization (1937), 1–6.
Das, S., Sun, X., Wang, F., Leboeuf, C., 2015. Estimating likelihood of future crashes for crash-prone drivers. Journal of Traffic and Transportation Engineering (English Edition) 2 (3), 145–157.
Eguakun, G., Park, P.Y., Quaye, K., 2015. Identifying optimal high-risk driver segments for safety messaging. Transportation Research Record: Journal of the Transportation Research Board 2019, 167–176.
Factor, R., Mahalel, D., Yair, G., 2008. Inter-group differences in road-traffic crash involvement. Accident Analysis and Prevention 40, 2000–2007.
Federal Motor Carrier Safety Administration (FMCSA), 2018. Pocket guide to large truck and bus statistics. United States Department of Transportation, Washington, D.C.
Girotto, E., Mesas, A.E., Andrade, S.M., Birolim, M.M., 2013. Psychoactive substance use by truck drivers: A systematic review. Occupational and Environmental Medicines 71 (1), 71–76.
Green, E.R., Ross, P.A., Blackden, C.L., Police, K.S., 2017. Kentucky traffic collision facts 2016. Kentucky Transportation Center Research Report.
Hasselberg, M., Vaeza, M., 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 for Highway Safety (IIHS), 2017. Fatality facts 2017. Insurance Institute of Highway Safety. Highway Loss Data Institute, Washington D.C.
Insurance Institute for Highway Safety (IIHS), 2018. Fatality facts 2018 - large trucks. Insurance Institute for Highway Safety, Highway Loss Data Institute, Washington DC.
Kocatepe, A., Ulak, M.B., Ozguven, E.E., Horner, M.W., 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.
Laerd Statistics, 2018. Point-biserial correlation using spss statistics.
Lambert, J., Gong, L., Elliott, C.F., Thompson, K., Stromberg, A., 2018. rFSA: An r package for finding best subsets and interaction. The R Journal 10, 295–308.
Lee, J., Abdel-Aty, M., Choi, K., 2014. Analysis of residence characteristics of at-fault drivers in traffic crashes. Safety Science 68 (0), 6–13.
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., 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.
Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road casualties in England. Accident Analysis and Prevention 973–984.
Ntsb, 1990. Fatigue, alcohol, other drugs and medical factors in fatal-to-the-driver heavy truck crashes 1 National Transportation Safety Board, U.S. Department of Transportation, Washington, DC.
PennState, 2018. Analysis of discrete data: Logistic regression. Elberly College of Science.
Souza, J.C., Paiva, T., Reimão, R., 2005. Sleep habits, sleepiness and accident among truck drivers. Arquivos de Neuro-Psiquiatria 63 (4), 925–930.
Stamatiadis, N., Deacon, J., 1997. Quasi-induced exposure: Methodology and insight. Accident analysis and prevention 29, 37–52.
Stamatiadis, N., Puccini, G., 1999. Fatal crash rates in the southeastern United States: Why are they higher? Transportation Research Board (1665), 118–124.
U.S. Census Bureau, 2016. American census survey.
World health Organization (WHO), 2018. Road traffic injuries. World Health Organization.
Zephaniah Jr., S., Smith, S.J., Weber, R.J., 2018. Spatial dependence among socio-economic 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., Jones, M.P., 2005. Fatal motor vehicle crashes in rural and urban areas: Decomposing rates into contributing factors. Injury Prevention 11 (1), 24–28.