Spring 4-15-2021

NIGHTTIME PEDESTRIAN FATALITIES ACROSS THE UNITED STATES AND THEIR RELATIONSHIP WITH BUILT ENVIRONMENT AND SOCIOECONOMIC FACTORS: GIS-BASED ANALYSIS AND ASSESSMENT

Amir Tarighi

Amir Tarighi

Candidate

Civil, Construction, & Environmental Engineering

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Dr. Nicholas Ferenchak , Chairperson

Dr. Joni Palmer

Dr. Haobing Liu

NIGHTTIME PEDESTRIAN FATALITIES ACROSS THE UNITED STATES AND THEIR RELATIONSHIP WITH BUILT ENVIRONMENT AND SOCIOECONOMIC FACTORS: GIS-BASED ANALYSIS AND ASSESSMENT

BY

AMIR TARIGHI

Bachelor of Science – Urban Planning & Engineering

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

Civil Engineering

The University of New Mexico
Albuquerque, New Mexico

May 2021

ACKNOWLEDGMENTS

I would like to sincerely thank my academic advisor and thesis chair, Dr. Nicholas Ferenchak, for his constant guidance, encouragement, support throughout the course of this research. He has taught me an immeasurable amount about academic research and his valuable feedback at every stage of this research was instrumental in the success of this thesis. I would also like to thank my second committee member, Dr. Joni Palmer, for her knowledge and expertise in research and urban planning concepts. Principles learned in her Urban System course and being her assistant in cultural planning research project significantly contributed to the foundation of this thesis. I would also like to thank my third committee member, Dr. Haobing Liu, for his help in evaluating my defense. Last but not least, I would like to thank my parents and my older brother for providing endless love and believing in me throughout my entire academic career.

NIGHTTIME PEDESTRIAN FATALITIES ACROSS THE UNITED STATES AND THEIR RELATIONSHIP WITH BUILT ENVIRONMENT AND SOCIOECONOMIC FACTORS: GIS-BASED ANALYSIS AND ASSESSMENT

By

AMIR TARIGHI

Bachelor of Science – Urban Planning & Engineering, July 2017
M.S., Civil Engineering, The University of New Mexico, May 2021

ABSTRACT

Fatalities from motor vehicle collisions are the fifth leading cause of years of lost life worldwide (Foreman et al., 2018) and are a great concern for all communities. This road safety crisis has been especially pronounced for U.S. pedestrians over the last decade. National records show that between 2010 and 2019, pedestrian fatalities in the U.S. increased 44% while all other traffic fatalities increased only 4% and more than fifty percent of pedestrian fatalities occurred in dark lighting conditions. This study analyzed the relationship between pedestrian fatality locations at night and five demographic characteristics of those locations including population density, median income, means of travel to work, educational attainment, and race at the census tract level from 2010 to 2019. A significant correlation between these demographic variables and the pedestrian fatality locations was discovered. Based on these demographic variables a predictive model has been built using the Generalized Estimation Equation approach that can be a great tool for planners and transportation agencies in identifying high-risk pedestrian fatality locations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Walking as a sustainable mode of transportation has been increasingly encouraged and studies have found that that there is a direct link between sustainability and community's walkability (Cubukcu, 2013). Sustainability is a complex and broad concept, and it has many indicators. In general, sustainable development is a development that considers the environmental health, social equity, and economic vitality needs of the present communities without compromising the ability of future generations to meet their own needs (UCLA Sustainability Committee, 2021). Based on this definition, the rate of walkability among the communities is one of the indicators of sustainability since walkable communities increase the active modes of traveling (e.g. walking and biking) among the residents and that decreases the personal vehicle miles traveled. Subsequently, it reduces greenhouse gases emission which eventually promotes health and many other sustainable development variables among the communities. However, in order to have walkable communities, pedestrian safety is an important factor to consider. Unfortunately, fatalities from motor vehicle collisions are the fifth leading cause of years of lost life worldwide (Foreman, et al., 2018) and are a great concern for all community officials, authorities, and any transportation planning agencies working on making a safe and reliable transportation network. This road safety crisis has been especially pronounced for U.S. pedestrians over the last decade. According to the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS) between 2010 and 2019, pedestrian fatalities in the U.S. increased 44% (Figure 1) while all other traffic fatalities increased only 4%.

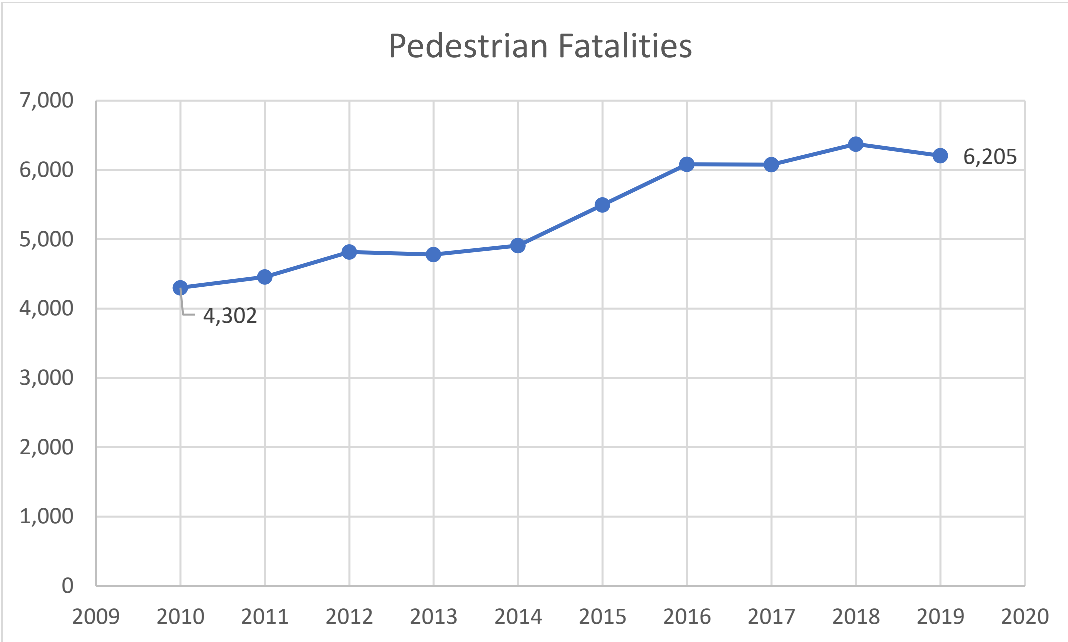

Figure 1. Pedestrian fatalities in U.S. (FARS)

Records from 2009 to 2018 show that more than fifty percent of pedestrian fatalities occurred in dark lighting conditions and while the fatalities in the daytime had been almost constant from 2009 to 2018, there was a dramatic increase in the occurrence of pedestrian fatalities in dark lighting conditions. Figure 2 was obtained from the pedestrian traffic fatalities report prepared for Governors Highway Safety Association (Richard Retting & Sam Schwartz Consulting, 2020). It shows how dramatically pedestrian fatalities in dark conditions have increased over time.

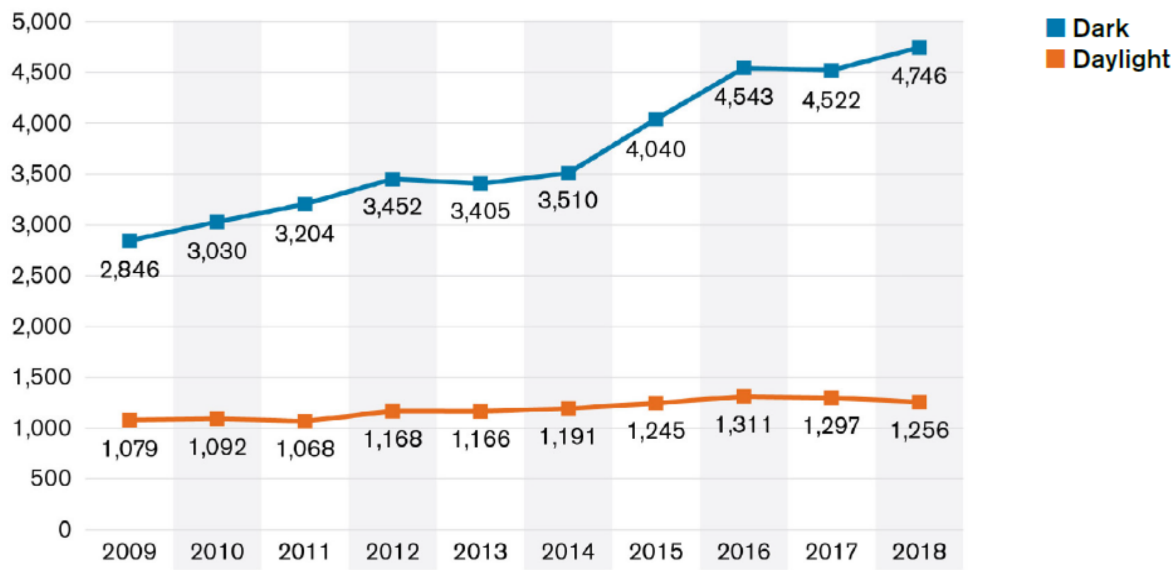

Figure 2. Pedestrian fatality trend by light condition

This research focuses on these nighttime vehicular crashes that led to the increase in pedestrian fatalities in the U.S. Specifically, I explored whether there is a relationship between fatality locations and built environment or socioeconomic characteristics of those locations.

I obtained motor vehicle fatality data from the Fatality Analysis Reporting System (FARS). This data accounts for all fatal collisions on public roadways where someone died within 30 days of the collision from 2010 to 2019. ArcGIS was used to georeference the coordinates of all fatal nighttime pedestrian collisions. Subsequently, census tract demographics including population density, median household income, mean of travel to work, race/ethnicity, and level of education were incorporated with the fatalities data. The data has been analyzed in R programming language to project a model for factors contributing to nighttime pedestrian fatalities and providing quantitative reasoning for the dramatic increase of such fatalities in recent years.

1.2. Goals & Objective

The results of this study will identify spatial factors correlated with the occurrence of nighttime pedestrian fatalities. Since transportation engineers and planners have limited resources to combat traffic safety issues, a better understanding of where to prioritize and invest these resources is an important pursuit. I have built a model for identifying high-risk pedestrian fatality locations which would be a great tool for planners and transportation agencies to implement traffic safety countermeasures in those areas. This study can widen transportation agencies' perspectives on factors associated with pedestrian fatalities and eventually help them to provide a safer transportation network system.

The thesis is organized into seven sections. The following is a brief explanation of each section.

- Section One: Introduction and an overview of the research process and research goals.
- Section Two: Literature reviews/ prior studies on the pedestrian fatalities demographic characteristics.
- Section Three: Data sources and processing
- Section Four: Methodological framework for Arc GIS Application and R programming language to find out the predictive model for pedestrian fatalities estimation.
- Section Five: Results and trends.
- Section Six: Discussion and interpretation of the results.
- Section Seven: Conclusion, study limitations, and future directions.

The next section reviews previous literature on this topic and supports the necessity for this study.

2. LITERATURE REVIEW

In recent years, walking as a sustainable mode of transportation has been increasingly encouraged in the sustainable development of societies. However, there are several barriers to achieving walkable communities. One of the most important factors is pedestrian safety. Safer neighborhoods contribute to a higher rate of walkability. One study showed that built environment and traffic safety barriers perceived by residents were one of the factors that contributed to the decline in the rate of walkability (Forman, et al., 2007). Another study showed that safety and traffic concerns were associated with predicting youth commute to schools (Kerr, et al., 2006). Unfortunately, fatalities from motor vehicle collisions were the fifth leading cause of years of lost life worldwide (Foreman, et al., 2018) and are a great concern for community officials. This road safety crisis has been especially pronounced for U.S. pedestrians over the last decade. According to FARS, between 2010 and 2019, pedestrian fatalities in the U.S. increased 44 percent and much of this increase occurred in dark lighting conditions. According to Pedestrian Traffic Fatalities by State (Richard Retting & Sam Schwartz Consulting, 2020), only about 21 percent of pedestrian fatalities in 2018 occurred in daylight.

Many factors play an important role in enhancing pedestrian safety. Previous research identified several different factors in pedestrian fatalities. Among them are driver and pedestrian behavior, lack of adequate roadway infrastructures (e.g., sidewalk, median, crosswalk, lighting condition), drug/alcohol involvement, and socioeconomic factors. Ferenchak and Abadi (2019) found that infrastructure and geometric design of roadways were primary factors in nighttime pedestrian fatalities. Drug and alcohol and consumption have been also recognized as an important factor in motor vehicle collisions in the U.S. (Brady & Li, 2014).

Also, research has found age and gender to be important factors in pedestrian injuries. According to the Traffic Safety Facts report (NHTSA, 2020), the percentage rate of children (under 14) fatalities in traffic crashes in 2018 was 17% of the total number of people killed in motor vehicle crashes, while the percentage of age group 15-65 and 65+ were 65 percent and 18 percent respectively. Besides, the highest rate of pedestrian fatality among all age groups was senior citizens (80+). That shows the vulnerability of children and senior citizens in traffic crash collisions. This fact is also consistent with another finding stating, that senior citizens (people older than 80 years old) had the highest risk of mortality. (Retting, Ferguson, & McCartt, 2003). Another study showed that in areas with a lower proportion of children, pedestrian collisions are reduced (LaScala, Gerber, & Gruenewald, 2000). That is maybe because of the low awareness of children to the surrounding environment. The gender composition of a population is another factor in pedestrian collisions. A study conducted in the UK shows that males had been more involved in pedestrian injuries (about 60%) based on data from 1993 to 2006 (Martin, 2006). In the U.S. also about 69% of pedestrian fatalities were male in 2018 (NHTSA, 2020).

Another important demographic contributing factor in pedestrian fatalities is the level of income. Studies show that pedestrian fatalities are more likely to happen in low-income census tracts (Rivara & Barber, 1985; Pharr, Coughenour, & Bungum, 2013; Ukkusuri, Hasan, & Abdul Aziz, 2011). Another study also found that pedestrian injuries increased in areas with a higher rate of poverty (Chakravarthy, Anderson, Ludlow, lotfipour, & Vaca, 2010). In this study, they used census tract demographic information and geocoded the crashes locations from 2000 to 2004 in a large Southern California County, and found out that a one percent increase in the percentage of residents with low income contributes to a 2.8 percent increase in pedestrian crashes. Other studies indicated the rate of unemployment as a factor, which is related to the level of median income (LaScala, Gerber, & Gruenewald, 2000; Ukkusuri, Hasan, & Abdul Aziz, 2011). Low car ownership was identified to be a contributing factor to pedestrian injuries as it leads to higher levels of active transportation and therefore more vulnerable road user exposure to traffic collisions (Abid, Leon-Rivera, Sartorio, Schrader, & Smith, 2019; Chimba, Musinguzi, & Kidando, 2018).

Race and ethnicity were also found as contributing factors in pedestrian collisions. In a city of Memphis study, in nonwhite census tracts, children pedestrian injuries rate with bigger household members were higher than adults (Rivara & Barber, 1985). Another pedestrian injuries and death study addressing the relationship between pedestrian injuries and race (in Arizona based on data from 1990-1996) indicated that American Indians had a significantly higher rate of pedestrian fatalities in all age groups than other races (Pharr, Coughenour, & Bungum, 2013). Another study showed that in New York from 2002 to 2006 the rate of pedestrian fatalities was higher in census tracts with a high number of Hispanics in the population (Ukkusuri, Hasan, & Abdul Aziz, 2011).

Level of education is also considered an important contributing factor in pedestrian crashes. A study of Hispanic children pedestrian injuries showed that parent's inability to read is related to pedestrian injuries, among other social and cultural factors (Agran, Winn, Anderson, & Del Valle, 1998). Another study in the city of San Francisco, CA found out that pedestrian injuries are most likely to happen in areas with a low level of education (LaScala, Gerber, & Gruenewald, 2000). A New York study based on the crash records from 2002 to 2006, revealed that there is a strong correlation between the uneducated population and pedestrian collisions (Ukkusuri, Hasan, & Abdul Aziz, 2011).

Additionally, urban sprawl and population density are interrelated and both have been found to be an important factor in pedestrian crashes frequencies. In a study of 448 US counties, researchers found that for each percent increase in the urban sprawl index, the pedestrian fatality rates were reduced from 1.47 percent to 3.56 percent (Ewing, Schieber, & Zegeer, 2003). Another study exploring the association between pedestrian crashes and socio-economic variables in Las Vegas, NV, revealed that population density and pedestrian crash frequencies have a reverse relationship (Pharr, Coughenour, & Bungum, 2013).

Moreover, land use characteristics of the built environment have been identified in many studies as an important factor of pedestrian fatalities. Population density is significantly affected by land use type and as mentioned earlier, population density is a contributing factor in pedestrian injuries. One study found a positive relationship between commercial land use areas and pedestrian collision frequencies (Osama & Sayed, 2017). However, since it is nearly impossible to gather land-use data on a national level, I used population density as a proxy.

According to NHTSA and FARS data, more than 50 percent of the pedestrian fatalities happened in dark lighting conditions. Only in 2018, about 76 percent of total pedestrian fatalities occurred at dark lighting conditions (NHTSA, 2020). This fact is also aligned with other studies conducted in the UK indicating that darkness would increase the chance of a collision by four times (Martin, 2006). Based on this reason, this study is focused on pedestrian fatalities that occur in dark lighting conditions.

Many studies tried to model and predict the crash frequencies using different methods of statistical modeling. Different regression models have been used widely. One of the regression analysis methods that has been found to be a good representation of longitudinal data is the

Generalized Estimating Equation (GEE) (Lord & Persaud, 2000; Hanley, Negassa, Edwardes, & Forrester, 2003; Ballinger, 2004). GEE approach was introduced by Liang and Zeger in 1986 to produce more efficient regression analysis of correlated observations in longitudinal data or clustered data. This method has been widely used in medical and life sciences (e.g., biology) as they frequently study repeated measurements over a long period of time. In this study, the data have been collected from the time period of 2010 to 2018. Thus, due to the longitudinal data utilized in this study, the GEE model has been chosen to model pedestrian fatalities.

Most of the previous studies' scale was limited to the city level and county level. Besides, less attention was on night nighttime pedestrian injuries and death. This research aims to provide a model by using socio-economic variables in pedestrian fatalities at night to predict the pedestrian fatalities at nighttime.

3. DATA

I obtained pedestrian fatalities records from 2010 to 2019 from FARS as excel spreadsheets. FARS is a nationwide data collection system created by the NHTSA in order to provide censuses on police reports of motor vehicle crashes contributing to the fatality that occur on public roadways. The death, however, must occur with 30 days in the United States (National Center for Statistics and Analysis, 2021). The FARS data provide detailed information about all people and vehicles involved in each collision including time, year, location coordinates, alcohol consumption, weather condition, the severity of the injury, and so on. These data can be used for a variety of purposes. One of them is to identify any possible trend in crashes fatalities as well as identifying hot spot crash locations. In general, the FARS data is widely used as a complementary information source for safety programs in the U.S. that are intended to enhance a safe and reliable transportation system (NHTSA, 2010).

For this study, motor vehicle crashes records from FARS data that involve pedestrian fatalities have been selected. Also, as this study aims to study nighttime pedestrian fatalities, the time of collisions has been filtered only to the dark condition. For doing that, the light condition column only chosen to dark time including lighted, not lighted, and unknown. Lighted or not lighted refers to the streetlight condition. For each year from 2010 to 2019, the same process has been done to provide only pedestrian fatality records at nighttime. FARS data include every incident coordination that we have used to geolocated locations in ARC GIS environment.

The census tract demographic information for each year has been obtained from National Historical Geographic Information System (NHGIS). This is a data system that contains longitudinal census data, which is able to be applied and integrated with Geographic Information System (GIS). Thus, NHGIS data are a great tool for data analysis within the GIS platform and for doing a variety of spatial analyses. This system provides easy online access to time series tables that include census information and also geographical data free of charge. Besides, the geographical boundary files are stored in the system as shapefiles which can be easily imported in GIS. The census tract's boundary shapefile has not been changed throughout the study period, allowing for consistent analysis.

For each year from 2010 to 2019 the census tract demographic information including the total population, race, means of transportation to work, educational attainment for the population 25 years and over, and median household income in the past 12 months has been collected as excel spreadsheets. The data source is American Community Survey (ACS) which is conducted by U.S. Census Bureau. ACS provides valuable demographic information at different scales (including cities, counties, tracts, and block groups) that is widely used in planning and decision-making purposes.

The variable race has ten categories including Total, White alone, Black or African American alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and Other Pacific Islander alone, some other race alone, Two or more races, Two or more races (Two races including Some other race), and Two or more races (Two races excluding Some other race, and three or more races). For this study, the race categories have been summarized into two different categories of white alone and nonwhite since many studies found a significant gap in the rate of pedestrian fatalities among these two categories (Udell, Daley, Johnson, & Tolley, 2014). For calculating the number of nonwhite population for each census, the number of white alone race has been subtracted from the total number of all races. The numbers then have been converted to the percentages for each census tract.

The means of travel to work also has several categories that have been sorted into two categories: walking to work and non-walking to work. In order to find out the number of people not walking to work, the number of total people walking to work has been subtracted from the total number of people travel to work for each census tract. The numbers then have been converted to the percentages for each census tract.

Another factor included in the calculation is the level of education. Educational attainment for the population 25 years and over has been categorized into two categories: number of bachelor’s holders and above degree and number of people holding below bachelor’s degree. The numbers then have been converted to the percentages for each census tract.

The population density has been obtained by dividing a census tract's total population by the census tract's geographic area. The geographic area is provided automatically by ARC GIS once the census boundaries shapefile imported to the GIS. For each year from 2010 to 2019, census boundaries have been obtained from the NHGIS as a shapefile.

4. METHODOLOGY

In this section, a detailed description of the methodology for making a predictive model of pedestrian fatality using census tracts' demographic information will be explained.

The first step for this research was to geolocate the pedestrian fatality locations in GIS. As mentioned in the previous section, for each year from 2010 to 2019 an excel table containing pedestrian fatalities information has been obtained through the FARS and filtered only by dark time. Each pedestrian fatality table has coordinates in the form of latitude and longitude which describe a point on the earth's surface. In order to show the fatalities coordinates which are in the form of x,y coordinates in GIS as a feature layer, the XY Point Data option in Add data menu has been used. This option will add a feature layer of pedestrian fatality points to the current map with the user-defined coordination system. For permanently saving the pedestrian fatalities feature layer, the layer has been exported and saved to a shapefile. Besides, for correct displaying of the fatalities points, the proper data frame coordination system has been defined in the Arc GIS environment. In order to obtain the best accurate locations and measurements of fatalities in the United States, the North American datum of 1983 (NAD 83) has been selected due to its suitability in fitting the earth surface in the North American region.

Once the pedestrian fatalities are shown in Arc GIS as points, they can be integrated into the census tracts. The census tract boundaries have been obtained through the NHGIS as a shapefile. Census tracts are geographic entities defined by the U.S. Census Bureau for population analysis. Census tracts usually contain a population size of 2500 to 8000.

For adding the census tract shapefile, the add menu option has been used. Once the census tract is added to the GIS, it will automatically be displayed using the previously defined coordinate system (which is NAD 83 in this study).

The last thing that added to the GIS environment before starting the spatial analysis process, is adding the demographic tables corresponding to each census tract. As described in section three, five demographic variables of population density, race, educational attainment, means of travel to work, and median income have been obtained through the NHGIS for all census tracts. After adding all the required files and tables, they have to be integrated to provide a shapefile containing valuable information for further analysis.

The important key is to have the same column with identical values for each census tract for both files that are intended to join. The NHGIS has already provided a column named “GISJOIN” which contains an identical value for each census tract. This column exists in both the table file (containing demographic information of census tracts) and the census tract boundaries shapefile.

After joining the demographic information to each corresponding census tract, the fatality points were spatially joined to find out how many fatalities occurred in each census tract each year.

The join and relate menu in GIS has an option for joining data from another layer based on the spatial location. Here the fatality points data has been spatially joined to the census tract shapefile. In order to get the number of fatality incidents that occurred in each census tract, there is an option that can provide a summary of the numeric attributes of the points that fill inside it and by checking the count option, a count field will be added; showing how many fatality points fall inside each census tract polygon.

After spatially joining the pedestrian fatalities incident points to the census tract polygons, there was an informative attribute table for census tracts shapefile containing demographic information and the number of pedestrian fatalities that occurred that year. This attribute table has been then, exported to excel files for further analysis.

Besides, a summary table for each year containing the average value of mentioned demographic variables for all the U.S. census tracts from 2010 to 2019 has been provided. These data were ready to be evaluated and finally used to build a model. As the different demographic variables' values, scale and metrics were different, all the demographic variables were converted to standardized values (also called z-scores). The Z score gives a numerical measurement for comparison by indicating how far a value is from the mean in a group of population. In other words, z-score helps analysts compare scores that come from different data set, to one another more accurately. In this study, the z-score enabled me to compare demographic values that had different normal distributions and metrics. To calculate the standardized value for each variable, the following formula has been applied.

$$Z = \frac{X\ (observation) - \mu(mean)}{\delta(standard\ deviation)}$$

Each year table first has been compiled to make a one spreadsheet containing all pedestrian fatalities during that ten years (2010 to 2019) with each census demographic information. Then the file has been imported to the R studio; a great platform to apply R programming language. R is a programming language designed for computing a variety of statistical analysis, handling data analysis, and graphics (What is R?, n.d.).

Importing and reading the CSV tables has done with the read.csv command and subsequently, the table columns have been converted to the numerical character to be able to perform analytical analysis. After importing the census tracts table that at least has one pedestrian fatality from 2010 to 2019, the mean of demographic variables values has been taken by running the following code and then has been graphically visualized in a table with a "formattable" package. This package is designed to create data frames in the format of HTML. All the U.S. census tracts' demographic variables mean from 2010 to 2019 are also imported to R studio in order to compare the trend among the U.S census tracts and census tracts that contains at least one pedestrian fatality.

In order to graphically show the trend and compare the results of all census tracts demographic variable mean with the demographic mean of the census tract that has at least one pedestrian fatality, the ggplot package has been used. Ggplot package is dedicated to creating complex plots from data in data frames.

The demographic variables showed a significant correlation, which can be used to build a prediction model for pedestrian fatality. The method used for modeling is Generalized Estimating Equations (GEE). GEE models are a good representation of longitudinal data specifically if they are counts data (Lord & Persaud, 2000; Hanley, Negassa, Edwardes, & Forrester, 2003; Ballinger, 2004). In this study, the data have been collected from the period of 2010 to 2019. Due to the longitudinal data type of this study, the GEE model has been used to model pedestrian fatalities. However, before generating the model it is important to find out if the variables are correlated to each other or not. This is called multicollinearity. Multicollinearity is the situation in a regression model in which two independent variables are highly correlated (Alin, 2010). Multicollinearity causes unreliability in models and can adversely affect the model results. For identifying the multicollinearity between variables first the correlation among all pairs of variables has been calculated. The common type of correlation that is usually used in data analysis is Pearson's Correlation that identifies two variables' possible linear association when both variables are normally distributed (Mukaka, 2012). The correlation coefficient value varies from -1 to 1 which greater absolute value determines a stronger relationship among the variables. There is a command in R for identifying the correlation value. Table 1 shows the summary of all pairs of the demographic variables' correlation values.

	Income	Population Density	White Percentage	Walking to Work Percentage	Bachelor and Above Degree Percentage
Income	1				
Population Density	-0.0351556	1			
White Percentage	0.2642081	-0.3033655	1		
Walking to Work Percentage	-0.1542899	0.2889286	-0.07511971	1	
Bachelor and Above Degree Percentage	0.726856	0.08311201	0.2097324	0.114166	1

Table 1. Pearson's correlation value of demographic variables

From the table, it can be seen that the bachelor and above degree coefficient and the income correlation coefficient are relatively high and thus are highly correlated. We need to determine which variable to remove from our model.

The variance inflation factor (VIF) is another tool to identify multicollinearity in regression analysis (Craney & Surles, 2002). VIF measures the degree of "how much the variance of the estimated coefficients increases is due to collinear independent variables" (Craney & Surles, 2002) and it is calculated by taking an independent variable and regress it against other independent variables. This will give a new R square in which with using the below formula the VIF for each variable would be calculated. Generally, R square indicates the model's strength and measures to what extent the dependent variable variance is the result of its relationship with independent variables. The R square varies from 0 to 1 and values closer to 1 indicate a stronger model.

$$\text{VIF} = \frac{1}{1-r^2}$$

r = r square

Table 2 shows the VIF for each of the explanatory variables. The VIF value starts from 1 and has no limit. Studies show that VIF greater than 2.5 might cause problems.

Year	Population Density	White Percentage	Walking to Work Percentage	Bachelor and Above Degree Percentage	Income
1.0002	1.031966	1.027987	1.0366	1.52028	1.5537

Table 2. VIF value of demographic variables

Income and bachelor and above degree variables show higher VIF values compare to the other variables in the model. Thus, based on the Pearson Correlation and VIF for both variables, these two variables are highly correlated. This can adversely affect the model result. So, one of them that is less significant in the model must be removed. For doing that, the GEE model has been generated by the running GEE function code in R studio and found out the variable income is less significant and weaker than the bachelor and above degree variable. Table 3 shows the summary of the model result.

```
Coefficients:
              Estimate   Std.err     Wald Pr(>|W|)    
(Intercept) -1.236e+02 3.816e+00 1048.236  < 2e-16 ***
Year         5.982e-02 1.894e-03  997.564  < 2e-16 ***
PopDens_z   -1.792e-01 9.791e-03  334.862  < 2e-16 ***
White_z     -2.170e-01 4.925e-03 1942.207  < 2e-16 ***
Walk_z      -1.994e-02 6.448e-03    9.558  0.00199 ** 
Bachelor_z  -1.935e-01 9.344e-03  428.883  < 2e-16 ***
Income_z    -7.260e-02 8.963e-03   65.596 5.55e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3. Summary of the model result (1)

The Wald Chi-Squared Test shows how significant a variable in the model is. The income variable wald value (65.596) is smaller than the bachelor and above degree variable (428.883). So, the income variable has been removed from the model. Also, the percentage of people who walk to work (indicated by Walk_z in the table) is so small; thus, it has a little impact on the model which can be removed to improve the model result. The final model has been built after removing the income variable and percentage of people who walk to work.

The distribution family chosen for this analysis is the poisson distribution as this is the best fit for discrete data and counts. Basically, for each GEE model, the distribution of the outcome should be assessed. The Poisson distribution is often used when the dependent variable represents discrete values such as count data. In this case, as we want to predict the number of incidents, thus the Poisson distribution has been selected.

5. RESULTS

In this section, I will explain the trend of the demographic variables from 2010 to 2019. Then, the model results will be discussed.

As explained in the previous section, the data after GIS spatial analysis has been imported to the R studio for statistical analysis and visualization. The data are in the form of excel spreadsheets containing all U.S. census tracts along with associated five demographic variables and the number of pedestrian fatalities that occurred each year.

Table 4 shows the result after taking the average of the census tracts demographic variables in R studio containing at least one pedestrian fatality at night.

Year	Median income	Population density	White population percentage	Walking to work population percentage	Bachelor and above degree holder population percentage
2010	48473.63	0.00180	66.61	3.04	22.04
2011	49591.36	0.00170	66.53	3.02	22.36
2012	49591.36	0.00180	66.33	3.00	22.44
2013	49096.82	0.00190	65.80	2.95	22.72
2014	49794.13	0.00170	66.84	2.82	22.80
2015	49146.80	0.00160	65.36	2.90	22.96
2016	50881.77	0.00180	65.24	2.93	23.65
2017	52684.92	0.00160	64.84	2.72	23.97
2018	55390.98	0.00170	64.41	2.62	24.42
2019	57672.05	0.00170	64.68	2.52	25.02

Table 4. The average value of demographic variables in census tracts having at least one pedestrian fatality

All of the U.S. census tracts' demographic variables mean from 2010 to 2019 are also imported and shown in table 5.

Year	Median income	Population density	White population percentage	Walking to work population percentage	Bachelor and above degree holder population percentage
2010	55130.48	0.00198	73.05	3.17	26.92
2011	56081.50	0.00199	73.51	3.20	27.16
2012	56205.38	0.00200	73.57	3.18	27.39
2013	56295.86	0.00202	73.41	3.17	27.69
2014	56831.44	0.00204	73.22	3.14	28.08
2015	57302.24	0.00205	73.05	3.12	28.52
2016	58810.32	0.00206	72.47	3.07	28.85
2017	61206.51	0.00208	72.54	3.04	29.55
2018	63825.66	0.00207	72.25	3.00	30.08
2019	66529.93	0.00207	72.00	2.96	30.64

Table 5. The average value of demographic variables in all U.S census tracts

The following figures show the trend of demographic variables graphically and compare the results of all census tracts demographic variable mean with the demographic mean of the census tract that has at least one pedestrian fatality. In the figures, the regression line also has been integrated. The red line and dots indicating all census tracts within the U.S. borders and the blue line and dots indicating all census tracts containing at least one pedestrian fatalities.

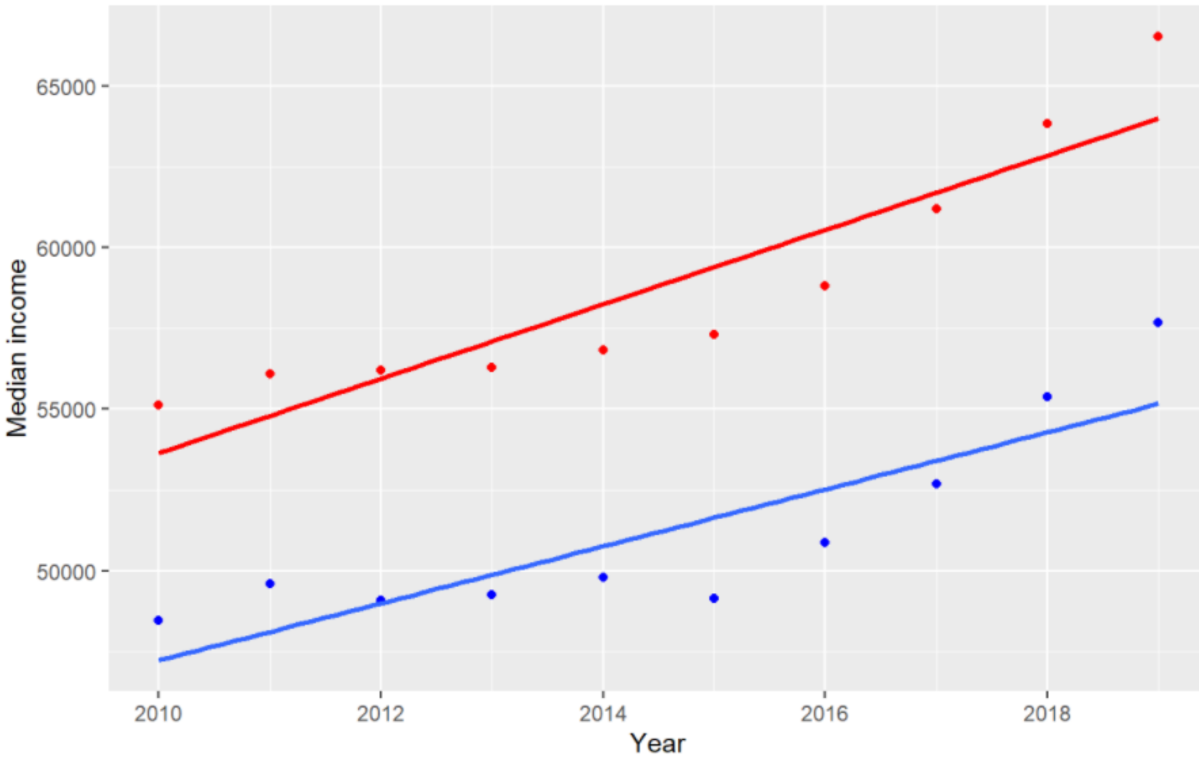

Figure 3. Median income trend

The above figure shows that median income in census tracts having at least one pedestrian fatality is less than the average of the U.S. census tracts. In other words, poor areas are more likely to be exposed to pedestrian fatality collisions than areas with a high level of residents income.

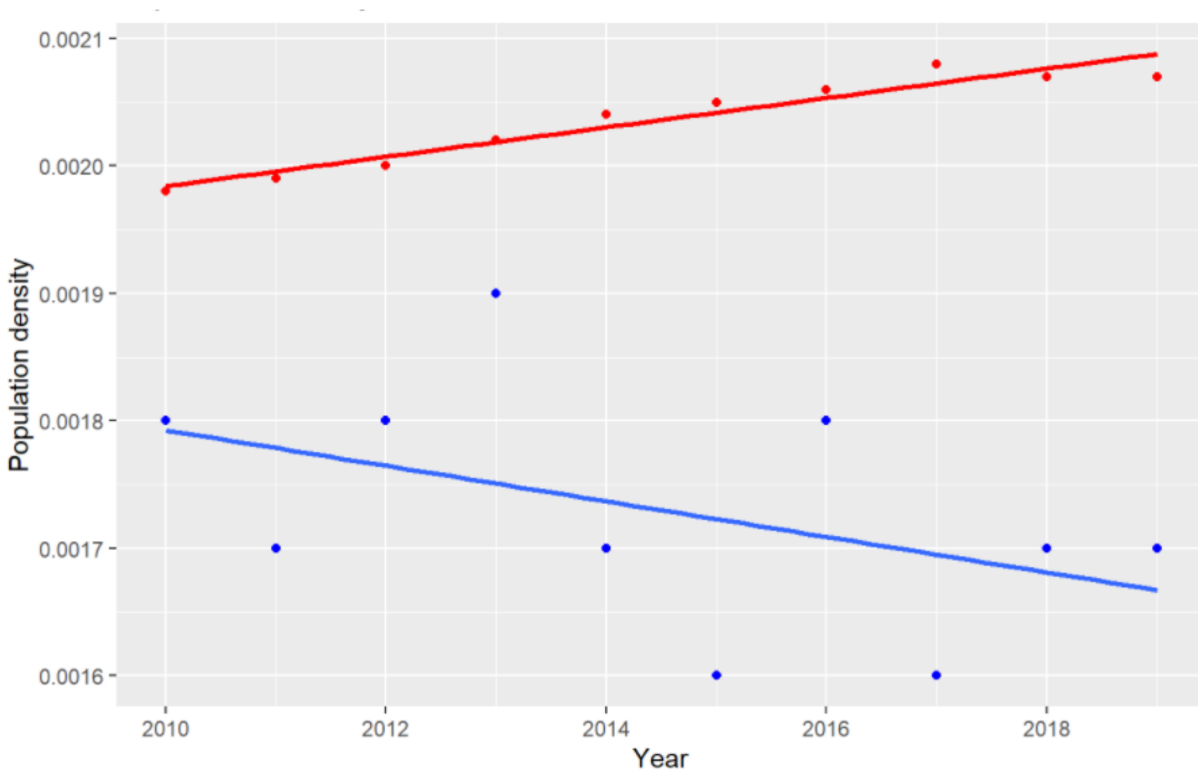

Figure 4. Population density trend

The population density has a different pattern in the census tract at least having one pedestrian fatality. That means as the population density increases in the U.S. over time, pedestrian fatalities were more likely to occur in places with low population density.

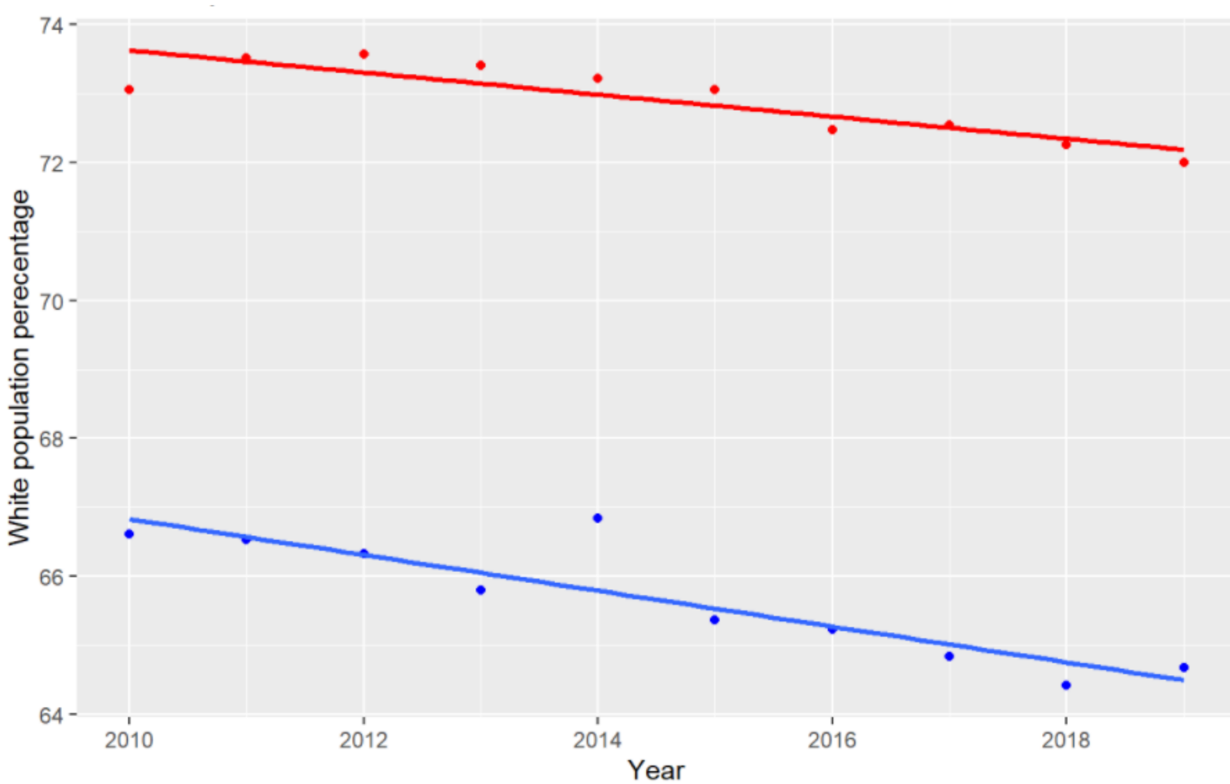

Figure 5. Race trend

The above figure shows that the white population percentage in census tracts having at least one pedestrian fatality is less than average of the U.S. census tracts and they both follow a similar pattern by the passage of time. In other words, in census tracts with a different race composition than white, the pedestrian fatality occurrence is higher.

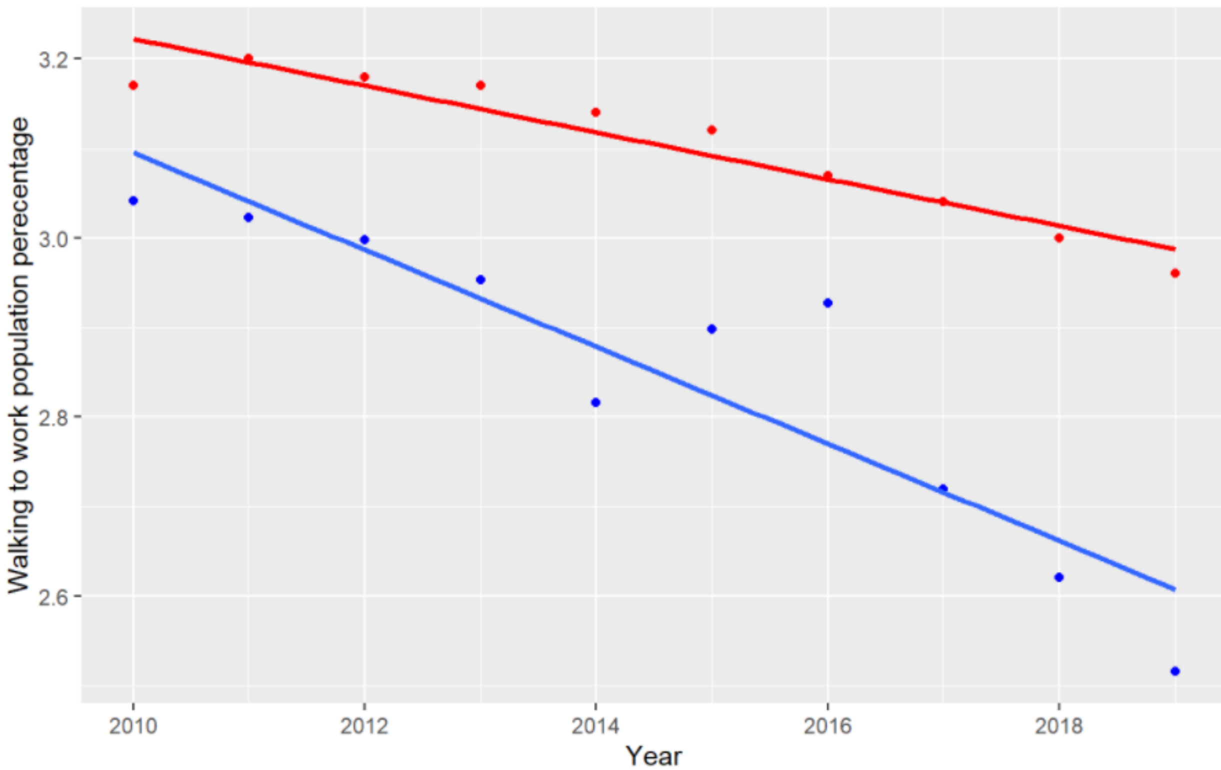

Figure 6. Means of travel to work (average percentage of people who walk to work) trend

The above figure shows that the percentage of the population who walk to work in census tracts having at least one pedestrian fatality is less than the U.S. average. That means in census tracts that people use walking as a means of traveling to work, have fewer pedestrian fatalities.

Figure 7. Level of educational attainment trend

The figure above shows that census tracts having a lower level of education are more exposed to pedestrian fatality.

As discussed in the previous section, the method used for building a predictive model of pedestrian fatality at night is Generalized Equation Estimation (GEE). However, before generating the model two variables of median income and means of travel to work have been removed. The variables median income and educational attainment level were highly correlated with each other that can adversely impact the model accuracy. Thus, median income variable has been removed as showing less significance than educational attainment variable in the model. The variable means of travel to work showed also weak significance in the model.

I have generated the model in R studio. The function for model generation is “Geeglm” and it builds a GEE model based on explanatory variables. The summary function provides result summaries of the GEE model (shown in figure 8).

```
##  Coefficients:
##              Estimate   Std.err Wald Pr(>|W|)
## (Intercept) -1.21e+02  3.80e+00 1016   <2e-16 ***
## Year         5.85e-02  1.88e-03  966   <2e-16 ***
## PopDens_z   -1.83e-01  9.70e-03  354   <2e-16 ***
## White_z     -2.26e-01  4.79e-03 2220   <2e-16 ***
## Bachelor_z  -2.46e-01  6.42e-03 1472   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
```

Figure 8. Summary of model results

The column “Estimate” in the above figure shows the model coefficients of independent variables (also called slope) (Nachtsheim, Kutner, Li, Neter, & Kutner, 2004) It measures how much of the dependent variable difference is explained by one unit of difference in the corresponding independent variable by one unit. For example, for the year variable, a one-unit increase would lead to a 0.0585 decrease in the number of pedestrian fatalities, meaning that nighttime pedestrian fatalities have been increasing throughout the study period.

All of the variables that were applied in the model have a significant code of 0 (***) indicating that all variables are significant in predicting pedestrian fatality.

The Wald Chi-Squared Test (Shown as “Wald” in the table) shows how significant is a variable in the model.

The final model indicated that race and level of education have the highest wald values respectively, thus indicating the most significant variables in this model in pedestrian fatality occurrence. Both of these variables follow a similar trend to the national trends and in this study, they were found as the most important factors in pedestrian fatalities. The variable population density had a different pattern than the national trend and in our model is not a quite strong variable for pedestrian fatalities.

5. DISCUSSION

This study analyzed the relationship between pedestrian fatality locations and the socioeconomic characteristics of those locations. Five demographic variables of population density, median income, means of travel to work, educational attainment, and race have been selected and analyzed at the census tract level from 2010 to 2019. While many pedestrian safety studies focused on small-scale geographic areas, this study scale contains all the U.S. census tracts.

In this study, I found a significant correlation between these demographic variables and the pedestrian fatalities locations. The result of this study showed that in census tracts with lower population density, median income, level of education, and percentage of people who do not walk to work, the rate of pedestrian fatality at night is higher. Population density, however, had a different pattern compared to other variables. Most of these variables follow overall demographic trends of the U.S., while the population density trend deviates from the national trend and showed that unlike the increase in U.S. population density, pedestrian fatality occurrence is more likely in lower population density areas. As previously mentioned, I recognized that in areas where the percentage of people who tend to walk to work is fewer, the pedestrian fatality rate is higher. Other studies indicated that dense areas can significantly increase the active modes of transportation among the residents, particularly walking (Udell, Daley, Johnson, & Tolley, 2014; Berkovitz, 2001). Thus, we can conclude that population density and means of travel to work variables are interrelated. The median income and educational attainment were also found to be significant factors in pedestrian fatality locations. I also found them to be highly correlated to each other through statistical methods. That makes sense as in areas with a higher level of education, people's chances to find high-paid jobs are higher. Both of these variables indicate that in areas with lower median income and level of education the rate of pedestrian fatalities is higher. That finding implicates the necessity for paying more attention to poor areas when taking traffic safety countermeasures. The variable of race was also found to be another significant factor in the pedestrian fatality locations in this study. The result showed that in areas with a lower rate of white population, the rate of pedestrian fatality is higher. This finding is in line with other findings indicating that pedestrian fatality was higher among the non-white populations (Pharr, Coughenour, & Bungum, 2013; Ukkusuri, Hasan, & Abdul Aziz, 2011). National records indicate that whites are more likely to have a college degree and subsequently that is a key for financial well-being (Pew Research Center, Social & Demographic Trends, 2016). That shows how specific race inequality can affect different socioeconomic statuses including lower median income and level of education and thus make these groups vulnerable to traffic safety issues based on this study findings. One of the high-risk pedestrian locations is U.S. suburbs and surprisingly “more than half of the people living below the poverty” (Benediktsson, 2017) reside in suburb areas. Low median income group living in suburbs (as a result of moving out from gentrified areas in city centers) are relying more on other modes of transportation than a private car (e.g., walking, biking, and transit) since they cannot afford it. Lower population density, private car ownership, and level of median income in suburbs are the variables that were found to have a great impact on pedestrian safety. Therefore, planning policies that prevent urban sprawl development and gentrification from urban centers including proper land use planning and design, would contribute to a safer transportation system. Land use planning and design have a great impact on urban density, the general shape of transportation network, an area's traffic indexes including traffic production and attraction, roadway design and thus speed limit, and travel distance (Hummel, 2001). Mixed land use planning with high urban density and short street blocks would decrease the number of trips, travel distance, private car dependencies, and promote the use of active modes of transportation (e.g. transit, walking, biking); often resulting in lower vehicular speed and less crashes severities (Hummel, 2001; Berkovitz, 2001). Therefore, development patterns that consider mixed land use designs, prevent gentrification, and suburbanization would contribute to a safer transportation network system.

In this study, I have used the Generalized Estimating Equations approach to build a predictive model based on year, population density, level of educational attainment, and race variables. The GEE model suitability for longitudinal and count data analysis was the reason for choosing that approach for this study. The model can be used in predicting an area pedestrian fatality rate. It would be a great tool for planners and transportation agencies, to predict high-risk areas. Subsequently, safety actions can be implemented in those areas.

6. CONCLUSION

Fatalities from motor vehicle collisions are among the top five leading causes of death each year in the U.S. (Centers for Disease Control and Prevention, n.d.). This road safety crisis has been especially pronounced for U.S. pedestrians. National records showed a dramatic increase in pedestrian fatalities (specifically at night) over the last decade. Although there are many risk factors associated with pedestrian collisions, this study investigated five socio-economic characteristics of pedestrian fatality locations from 2010 to 2019. The demographic variables of population density, median income, level of educational attainment, race, and means of travel to work have been analyzed through these ten years, and based on these data a predictive model has been built using the Generalized Estimation Equation approach.

This study found that in census tracts with lower population density, median income, level of education, and percentage of people who do not walk to work, the rate of pedestrian fatality at night is higher. Additionally, in areas that the majority of the population race is non-white, the rate of pedestrian fatality is higher. The studied demographic variables in this research indicate that poor nonwhite areas with lower population density are more likely to have a higher rate of pedestrian fatalities in crash collisions at night. That implicates the necessity for planners and transportation agencies to pay more attention to those areas in terms of traffic safety actions. One of the traffic safety policies regarding socioeconomic characteristics of high-risk locations that might have a great influence on improving pedestrian safety is proper land use planning. The land-use type of an area directly impacts the socio-economic characteristics of an area including but not limited to the population, employment, level of income. It also affects an area's traffic indexes including traffic production and attraction, roadway design and thus speed limit, and travel distance. All of these variables shown in this study and other studies have a direct relationship to traffic safety. Thus, it is suggested that urban policies encouraging sustainable modes of transportation (which increases walkability) and applying mixed land use design (leading to a higher rate of population density, decrease travel distance, and increase walkability) would contribute to safer transportation network system.

However, the socio-economic condition of a location is not solely implicating associated risk factors for pedestrian fatality. Built environment characteristics have been proved as an important factor in high-risk pedestrian injury areas. Built environment safety policies including roadway infrastructure safety design improvement (e.g., adequate lighting, sidewalk, crosswalk), driver and pedestrian safety awareness and training, applying the latest safety technologies, etc. are important to be considered in those locations. Thus, traffic safety issues involve complex and different factors and therefore, traffic safety issues need a systematic safety approach that incorporates all aspects of safety factors including built environment to socio-demographic characteristics of high-risk areas.

This study scale is at the national level. Future studies might focus on smaller scales such as a county or even a corridor for identifying more detailed risk factors associated with pedestrian fatality. Also, the census data used in this study might be sometimes misleading especially in regard to race and ethnicity. This usually happens more in areas with a lower rate of illiteracy. The reasons are including but not limiting to false information provided by the surveyor, confusion in identifying race and ethnicity, and survey collection systematic errors.

The predictive model was built based on four variables of year, population density, level of educational attainment, and race. For having a more accurate model in predicting fatality, more demographic variables can be used. Moreover, the built environment variables (e.g. road speed limits and lighting conditions) can be useful in pedestrian fatality predictive models.

Current technology advances in Machine Learning (ML) and computer vision have a great potential in enhancing pedestrian safety. ML approaches have great abilities in sorting the relative importance of attributes in explaining pedestrian crashes. Studies show that ML-based models in crash prediction models are more efficient than statistical ones as a result of their capability in dealing with regression and classification problems and multivariate response models (Silva, Andrade, & Ferreira, 2021). Besides, smart devices and systems such as Intelligent driver support systems equipped with artificial intelligence and machine learning algorithms can enhance pedestrian safety as well. For my Ph.D. I am planning to study machine learning algorithms for pedestrian safety purposes that can be used in a wide variety of systems and platforms to enhance traffic safety and providing a safe environment for pedestrians.

REFERENCES

Abid, Z., Leon-Rivera, A., Sartorio, J., Schrader, R., & Smith, J. (2019). Socioeconomics, the Built Environment and Pedestrian Safety in Central Florida.

Agran, P. F., Winn, D. G., Anderson, C. L., & Del Valle, C. (1998). Family, social, and cultural factors in pedestrian injuries among Hispanic children. *Injury Prevention, 4*, 188–193.

Alin, A. (2010). Multicollinearity. *WIREs Comp Stat, 2*, 370–374. doi:10.1002/wics.84

Ballinger, G. A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods, 7*(2), 127-150. doi:10.1177/1094428104263672

Benediktsson, M. O. (2017). Beyond the Sidewalk: Pedestrian Risk and Material Mismatch in the American Suburbs. *Mobilities, 12*(1), 76-96. doi:10.1080/17450101.2015.1019748

Berkovitz, A. (2001). *Federal Highway Administration Research and Technology*. Retrieved from Federal Highway Administration: https://www.fhwa.dot.gov/publications/publicroads/01septoct/marriage.cfm

Brady, J. E., & Li, G. (2014, March). Trends in Alcohol and Other Drugs Detected in Fatally Injured Drivers in the United States, 1999–2010. *American Journal of Epidemiology, 179*(6), 692–699. doi:https://doi.org/10.1093/aje/kwt327

Centers for Disease Control and Prevention. (n.d.). *FastStats*. Retrieved from CDC: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Chakravarthy, B., Anderson, C. I., Ludlow, J., lotfipour, S., & Vaca, F. E. (2010). The Relationship of Pedestrian Injuries to Socioeconomic Characteristics in a Large Southern California County. *Traffic Injury Prevention, 11*, 508–513. doi:10.1080/15389588.2010.497546

Chimba, D., Musinguzi, A., & Kidando, E. (2018). Associating pedestrian crashes with demographic and socioeconomic factors. *Case Studies on Transport Policy, 6*, 11-16. Retrieved from https://doi.org/10.1016/j.cstp.2018.01.006

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering, 14*(3), 391-403. doi:10.1081/QEN-120001878

Cubukcu, E. (2013). Walking for Sustainable Living. *Procedia - Social and Behavioral Sciences*, 33-42. doi:10.1016/j.sbspro.2013.08.335

Ewing, R., Schieber, R. A., & Zegeer, C. V. (2003). Urban Sprawl as a Risk Factor in Motor Vehicle Occupant and Pedestrian Fatalities. *American Journal of Public Health, 93*, 1541-1545.

Foreman, K. J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., . . . Murray, C. J. (2018, November). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for

2016–40 for 195 countries and territories. *Lancet*, 2052-2090. doi:10.1016/S0140-6736(18)31694-5

Forman, H., Kerr, J., Norman, G. J., Saelens, B. E., Durant, N. H., Harris, S. K., & Sallis, J. F. (2007). Reliability and validity of destination-specific barriers to walking and cycling for youth. *Preventive Medicine 46*, 311-316. doi:10.1016/j.ypmed.2007.12.006

Hanley, J. A., Negassa, A., Edwardes, M. D., & Forrester, J. E. (2003). Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. *American Journal of Epidemiology, 154*, 364–375. doi:10.1093/aje/kwf215

Hummel, T. (2001). Land use planning in Safer Transportation Network Planning.

Kerr, J., Rosenberg, D., Sallis, J. F., Saelens, B. E., Frank, L. D., & Conway, T. L. (2006, April). Active commuting to school: Associations with environment and parental concerns. *Medicine and science in sports and exercise, 38*(4), 787–794. doi:10.1249/01.mss.0000210208.63565.73

LaScala, E. A., Gerber, D., & Gruenewald, P. J. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. *Accident Analysis and Prevention, 32*, 651–658.

Lord, D., & Persaud, B. N. (2000). Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. *Transportation Research Record 1717*, 102-108. doi:https://doi.org/10.3141/1717-13

Martin, A. (2006). *Factors Influencing Pedestrian Safety: A Literature review.*

Mukaka, M. (2012). Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69-71.

Nachtsheim, C. J., Kutner, M. H., Li, W., Neter, J., & Kutner, M. (2004). *Applied Linear Statistical Models (McGraw-Hill/Irwin Series Operations and Decision Sciences)* (5th ed.).

National Center for Statistics and Analysis. (2021). *Fatality Analysis Reporting System (FARS) analytical user's manual, 1975-2019 (Report No. DOT HS 813 023).* National Highway Traffic Safety Administration.

NHTSA. (2010). *Report to CongressNHTSA's Crash Data Collection Programs.* National Highway Traffic Safety Administration.

NHTSA. (2020). *Traffic Safety Facts.* National Highway Traffic Safety Administration.

Osama, A., & Sayed, T. (2017). Macro-Spatial Approach for Evaluating the Impact of Socio-Economics, Land Use, Built Environment and Road Facility on Pedestrian Safety. *Canadian Journal of Civil Engineering*.

Pew Research Center, Social & Demographic Trends. (2016, June). Retrieved from Pew Research Cente: https://www.pewresearch.org/social-trends/2016/06/27/1-demographic-trends-and-economic-well-being/

Pharr, J., Coughenour, C., & Bungum, T. (2013). Environmental, Human and Socioeconomic Characteristics of Pedestrian Injury and Death in Las Vegas, NV. *International Journal of Sciences*(10).

Retting, R. A., Ferguson, S. A., & McCartt, A. T. (2003, September). A Review of Evidence-Based Traffic Engineering Measures Designed to Reduce Pedestrian–Motor Vehicle Crashes. *Am J Public Health, 93*(9), 1456–1463. doi:https://doi.org/10.2105/ajph.93.9.1456

Richard Retting, & Sam Schwartz Consulting. (2020). *Pedestrian Traffic Fatalities by State.* Governors Highway Safety Association. Retrieved from https://www.ghsa.org/resources/Pedestrians20

Rivara, F. P., & Barber, M. (1985). Demographic Analysis of Childhood Pedestrian Injuries.

Silva, P. B., Andrade, M., & Ferreira, S. (2021). Machine learning applied to road safety modeling: A systematic literature review. *Journal Of Traffic And Transportation Engineering (English Edition), 7*(6), 775-790. Retrieved from https://doi.org/10.1016/j.jtte.2020.07.004

UCLA Sustainability Committee. (2021). *Sustainability at UCLA*. Retrieved from UCLA Sustainability: https://www.sustain.ucla.edu/what-is-sustainability/

Udell, T., Daley, M., Johnson, B., & Tolley, R. (2014). Does density matter? The role of density in creating walkable neighbourhoods. *National Heart Foundation of Australia*.

Ukkusuri, S., Hasan, S., & Abdul Aziz, H. (2011). Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. *Journal of the Transportation Research Board, No. 2237*, 98–106. doi:10.3141/2237-11

What is R? (n.d.). Retrieved from The R Project for Statistical Computing: https://www.r-project.org/about.html

WHO. (2013). *Pedestrian safety: A road safety manual for decision-makers and practitioners.* World Health Organization.