2005

Spatial analysis of pedestrian accidents

Tomoyuki Hashimoto

University of South Florida

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the American Studies Commons

Scholar Commons Citation
http://scholarcommons.usf.edu/etd/2920

This Thesis is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Spatial Analysis of Pedestrian Accidents

by

Tomoyuki Hashimoto

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science
Department of Environmental Science and Policy
College of Arts and Science
University of South Florida

Co-Major Professor: Paul Zandbergen, Ph.D
Co-Major Professor: L. Donald Duke, Ph.D
Jayajit Chakraborty, Ph.D

Date of Approval:
January 27, 2005

Keywords: GIS, Hillsborough, land use, negative binomial, traffic safety

© Copyright 2005, Tomoyuki Hashimoto
DEDICATION

I dedicate this work to my parents, my idols, and to my family. Thank you for all the support you have given me throughout my life, for your care, and for the invaluable education I have acquired from you.
ACKNOWLEDGEMENTS

I would like to thank the ESP personnel for making my stay at USF memorable. I extend special gratitude to Dr. Paul Zandbergen for his helpful assistance in this research. Thank you to Dr. L. Donald Duke and Dr. Jayajit Chakraborty for their constant guidance and commitment to quality research.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2. Objective</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 2: LITERATURE REVIEW</td>
<td>5</td>
</tr>
<tr>
<td>2.1. Demographic Factors</td>
<td>5</td>
</tr>
<tr>
<td>2.2. Land Use Factors</td>
<td>9</td>
</tr>
<tr>
<td>2.3. Roadway Factors</td>
<td>13</td>
</tr>
<tr>
<td>2.4. Traffic Factors</td>
<td>14</td>
</tr>
<tr>
<td>2.5. Statistical Properties of Models</td>
<td>15</td>
</tr>
<tr>
<td>2.5.1. Multiple Linear Regression Model</td>
<td>15</td>
</tr>
<tr>
<td>2.5.2. Poisson Regression Model</td>
<td>16</td>
</tr>
<tr>
<td>2.5.3. Negative Binomial Regression Model</td>
<td>17</td>
</tr>
<tr>
<td>2.6. Spatial Analysis Techniques</td>
<td>20</td>
</tr>
<tr>
<td>2.7. Summary of the Literature Review</td>
<td>21</td>
</tr>
<tr>
<td>CHAPTER 3: METHODOLOGY</td>
<td>23</td>
</tr>
<tr>
<td>3.1. Data Collection</td>
<td>23</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1 Statistical Properties of Variables................................................................. 46
Table 2 Result of Negative Binomial Regression......................................................... 48
Table 3 Coefficient of Correlations for Independent Variables................................... 49
Table 4 Regression Result without Children Resident Rate....................................... 52
Table 5 Regression Result with Vulnerable Resident Rate ....................................... 53
Table 6 Negative Binomial Model with Population Density....................................... 55
Table 7 Negative Binomial Model with Residential Land Use Percentage.................. 56
Table 8 Regression Results for Block Groups with Traffic Count Data Inside............. 61
LIST OF FIGURES

Figure 1 Hillsborough County ................................................................. 2
Figure 2 Accident Point Data................................................................. 24
Figure 3 State Roadways ...................................................................... 25
Figure 4 Boundaries of Census Block Groups......................................... 27
Figure 5 Land Use Map ...................................................................... 29
Figure 6 Excluded Census Block Groups .............................................. 31
Figure 7 Accident Density Map ............................................................. 34
Figure 8 Four Land Use Types ............................................................... 35
Figure 9 AADT Points ....................................................................... 38
Figure 10 Distribution of The Dependent Variable ............................... 40
Figure 11 Distribution with A Radius of 250m ....................................... 41
Figure 12 Distribution with A Radius of 1000m ..................................... 42
Figure 13 Census Block Groups with Traffic Count Data ....................... 60
SPATIAL ANALYSIS OF PEDESTRIAN ACCIDENTS

Tomoyuki Hashimoto

ABSTRACT

Improving traffic safety for pedestrians is an urgent task for the Hillsborough County. The objective of this research is to understand the contributing factors that cause pedestrian accidents in Hillsborough County in the State of Florida. Specifically, this research attempts to determine the effects of demographic, land use, roadway and traffic volume factors on the number of pedestrian accidents. Five hypotheses were proposed to examine these factors. GIS was used to perform spatial analysis. Based on pedestrian accident records from 1999 to 2001, accident density map was created using Kernel density method. The value of the accident density was assigned to each census block group, and this value formed a dependent variable of the model. Accident models were developed using negative binomial regression to model the relationship between contributing factors and pedestrian accidents.

The results showed that commercial and service land use, residential land use, and the density of state roadways increase the number of pedestrian accidents. It was also found that average household income and the percentage of elderly residents lower the number of pedestrian accidents. The percentage of child residents did not show a statistically significant result. The effect of traffic volume was not determined because of data limitations. It should be explored in future research.
CHAPTER 1: INTRODUCTION

1.1 Background

In the United States, 4,995 people died in the year 2001 while walking down the street, up from the toll of 4,843 in 2000. An estimated 78,000 pedestrians were injured during each of those two years (Surface Transportation Policy Project, 2002).

The State of Florida has been suffering an enormous socio-economic loss due to pedestrian accidents. In the State of Florida, 16.8% of all traffic deaths were pedestrians, as compared to a national average of 12% in the year 2000-2001 (Surface Transportation Policy Project, 2002).

The study area of this research is Hillsborough County in the State of Florida (Figure 1). Hillsborough County is located midway along the west coast of Florida, with a population of 1,083,520 in 2003 (Metropolitan Planning Organization, 2004). The population is growing, and mass transit is not developed around this area. Severe demands are placed on the roadways. Also, for pedestrians, Tampa-St. Petersburg-Clearwater was ranked as the second most dangerous metropolitan area in the nation, based on an index of pedestrian deaths and percentage of commuters walking to work (Surface Transportation Policy Project, 2002). Improving safety for pedestrians is an urgent task for Hillsborough County.
1.2 Objective

The objective of this research is to understand the contributing factors that cause pedestrian accidents in Hillsborough County. Specifically, this research attempts to determine the effects of demographic, land use, roadway and traffic variables on the number of pedestrian accidents.

Five hypotheses are proposed to examine these factors. The first hypothesis is that residents in low income areas are more likely to experience pedestrian accidents because of their behavioral patterns (i.e. most of them don’t own cars and walk in the neighborhood more often compared to the residents in the other areas). The second hypothesis is that land use, in particular concentrated residential and commercial land use, have a significant effect on the number of pedestrian accidents because a large number of the people walk around these areas. The third hypothesis is that age distribution in an area affects the number of accidents. It is reasonable to assume children and elderly people are less careful when they face traffic and they are more likely to get involved in accidents. The fourth hypothesis is that the areas with high roadway density (the length of the state roadways divided by the total area) have more pedestrian accidents than other areas. The fifth hypothesis is that high traffic volume areas experience more accidents.

To examine these hypotheses, an accident model is developed using demographic, land use, roadway and traffic factors as independent variables. This research begins with a review of previous research that have attempted to relate traffic accidents (including pedestrian accidents) and contributing factors (Chapter 2). On the basis of this review, an appropriate modeling approach is selected to study the relationship between the
contributing factors and pedestrian accidents (Chapter 3). This is followed by model estimation and discussion of the model (Chapter 4). Finally, an overall summary of model findings and conclusions are provided (Chapter 5). If any of the independent variables have a statistically significant effect on the pedestrian accidents, and residents in particular areas have higher risk than in others, this will give useful information on transport and land use planning policies.
CHAPTER 2: LITERATURE REVIEW

The purpose of this review is to investigate the literature on the subject of pedestrian accident analysis. Four of the accident contributing factors are explored in this chapter, followed by a review of statistical models and spatial analysis techniques.

2.1 Demographic Factors

First, several studies have analyzed the effects of demographic factors on pedestrian accidents. Rivera et al. (1984) investigated the factors in a child’s living environment and socioeconomic background that contribute to the risk of pedestrian injury. 210 pedestrian injuries among children aged 0 to 14 years were reported in the city of Memphis in 1982. The City of Memphis had a population of around 650,000 at that time. For every census tract in the city, stepwise multiple regression analyses were performed to determine a combination of economic and environmental variables. Injuries per acre formed the dependent variable.

Census tracts with reported injuries had twice the percentage of non-white population, lower household incomes, more children living in female-headed house holds, more families living below the poverty level, and greater household crowding, compared with those without reported injuries.
Abdalla (1997) investigated the relationship between casualty frequencies and the distance of the accidents from the zones of residence. Abdalla’s research was conducted for the Lothian region in Scotland. While more driver, rider or passenger casualties occurred at longer distances (greater than 2,500 meters) from the zone of residence, the number of pedestrian casualties decreased with the distance from the zone of residence, and there was a high concentration of pedestrian casualties within 500 meters of the residence.

Abdalla compared the casualty rate in the most affluent areas (15%) with the most deprived areas (15%). In other words, the top 15% areas in average income and the bottom 15% areas. The rate in the most deprived areas per 10,000 residents was above the entire region’s rate, whereas in the most affluent areas it was below. Also, it was observed that for short distances (up to 500 meters) between the casualty’s residence zone and the accident location (not only pedestrian accidents), higher rates of casualties are seen in areas characterized as deprived rather than affluent.

Two important suggestions were provided by Abdalla’s research. First, pedestrian accidents are likely to occur within 500 meters of the residence. Second, residents in lower income areas are likely to show higher rates of pedestrian accidents.

Harruff (1998) performed a retrospective analysis of 217 pedestrian traffic fatalities in Seattle. All deaths classified as pedestrian traffic fatalities during six years (1990-1995) were analyzed. Harruff found that elderly pedestrians were most vulnerable because they were more likely to be injured as a pedestrian and more likely to die of injuries that a younger person might survive.
Petch et al. (2000) examined child road safety in the urban environment in the city of Salford (United Kingdom). Spatial distribution of the casualties was investigated using both statistical and spatial analysis techniques. As concluding remarks, it was stated that the majority of child pedestrian and cycle accidents occurred within 500 meters of the casualties’ place of residence.

LaScala (2000) explored some geographic correlates of pedestrian injury collisions through a spatial analysis of pedestrian accidents data. This study examined rates of pedestrian injuries across 149 census tracts in the city of San Francisco, which is a highly developed area. The number of motor vehicle collisions in which a pedestrian was injured or killed was aggregated within census tracts.

The results showed that pedestrian injury rates were related to population density, age composition of the local population, unemployment ratio, gender and education. Availability of alcohol through bars was directly related to pedestrian injury collisions in which the pedestrian had been drinking alcohol.

Graham (2003a) attempted to understand in detail how the nature of the local environment affected pedestrian casualties in England. He chose “wards” as the spatial unit for this analysis. The average area of “wards” was 14 square kilometers.

Graham adopted the number of pedestrian casualties in each ward as the dependent variable. Poisson regression and negative binomial regression modeling in case of over dispersion were used. It was stated that these models were appropriate because accident data were discrete in nature and had a preponderance of zeros and small values.
The result showed that the wards with a large population had higher incidence of pedestrian casualties than other wards. However, as population density increased, the incidence of accidents decreased. Also, the casualties increased with urban density, but this effect diminished for the most extremely dense wards due to the effect of urban congestion. Also, socio-economic status had a powerful negative influence on pedestrian casualties.

Graham (2003b) analyzed child pedestrian casualties in England, focusing on the influence of socio-economic deprivation. As well as the aforementioned study by Graham, 8414 wards were used as spatial analysis units. Using the same independent variables as the previous study, an accident model was developed for adult pedestrian accidents and children pedestrian accidents. Negative binomial regression modeling was used on the grounds that the discrete nature of account data and the preponderance of zeros and small values meant that the use of linear multiple regression could produce inefficient, inconsistent, and biased estimates. Also, overdispersion of the data convinced to use a negative binomial model instead of a Poisson regression model.

The result showed an association between increased deprivation and higher number of pedestrian casualties across England. The deprivation effect was strong both for all children casualties and for children killed or seriously injured. Estimates for adult casualties also revealed a positive and significant association with increasing deprivation, but the magnitude of the effect was smaller than for children casualties.

Noland et al (2003) explored statistical associations between road casualties and demographic variables for 8414 wards in England. To do so, statistical models were created to explain the incidence of accidents. “Wards” was chosen as the spatial unit for
analysis as well as Graham’s research. Poisson and negative binomial regression were used as opposed to multiple regression.

It was concluded that densely populated areas tend to have few traffic casualties while commercial areas have many traffic casualties. Also, the areas with higher levels of social deprivation had relatively higher casualty rates, but this effect was less strong when only motorized casualties were considered.

Noland and Quddus (2004) presented an analysis of pedestrian and bicycle casualties using cross-sectional time series data for 11 regions of the Great Britain. Regional data on pedestrian and bicycle road accidents over 20 years (from 1979 to 1998) were collected for the analysis. Negative binomial regression was used on the grounds that the distribution of accidents had a Poisson distribution with over-dispersion in the error term.

The findings included that increased expenditure on alcohol was positively associated with increased pedestrian and bicycle accidents. Also, analyses did not show a clear relationship between population age characteristics and the likelihood of casualties.

2.2. Land Use Factors

Some of the studies showed that certain land use types have an effect on the number of accidents. However, in some cases, actual land use maps were not available, so proxy variables were used to estimate the land use types.

Levine (1995a) examined spatial patterns in motor vehicle crashes on the Island of Oahu for 1990. Accidents were categorized as fatalities, serious injuries, alcohol-
related, any injuries, and all accidents. Also, accidents were examined for each hour in the 24-hr day and for weekdays and weekends.

The results showed that most accidents were closer to employment centers than to residential areas. In the suburban and rural areas, however, accidents were more likely to involve fatalities or serious injuries, and they were related to night-time driving and alcohol.

Levine’s follow-up study (1995b) attempted to explain the spatial patterns by population, employment and road characteristic variables in order to show that activities which generate trips also indirectly predict crashes. This method focused on characteristics of neighborhoods. In Levine’s research, the dependent variable was the number of accidents in 1990 in each census block group on the Island of Oahu in the state of Hawaii. Linear regression was used in the model.

The results showed that retail employment contributed the most to the number of accidents. It was suggested that two biases produced by assigning accidents to zones. First, accidents were assigned to zones, rather than specific locations, producing a spatial error. Second, the use of zones assumes that the risk of accidents was uniform at all locations within the zone, a situation which was frequently not correct. However, it was claimed that by choosing small and relatively homogenous zones, the advantages of grouping, namely the ability to associate characteristics of the zone with accidents, outweigh the biases produced by grouping.

Ng (2002) attempted to integrate the mapping and statistical techniques to develop a systematic algorithm to assess accident risk in Hong Kong, which is a highly developed area with a dense population.
The number of accident events in each Traffic Analysis Zone (274 zones) in Hong Kong was used as the dependent variable. Land use factors (27 types) were used as the independent variable on the grounds that they can influence human activities and human behavior, which affects the number of accidents. Ng attempted to explain accidents only by land use factors. Negative binomial regression was adopted as the statistical method. Ng stated that Negative binomial regression was found to be an appropriate form to mimic the relationship between the number of accidents and the land use factors.

There were important findings in Ng’s research. High-accident rate zones and low-accident rate zones showed different results. In high-accident rate zones, the number of cinema seats, commercial area, and flatted factory area were found to have positive and significant effects on the number of pedestrian-related accidents. The results also showed that greenbelt area, specialized factory area, and the number of territory school places had significantly negative effects.

However, in low-accident rate zones, there were no land use factors that had significant effects on the pedestrian-related accident occurrences. This result suggested that areas with high accident rates have different land use patterns compared to the areas with low rates. Also, it was stated that the algorithm was more efficient in the case of fatality and pedestrian-related accident analysis.

Kim and Yamashita (2002) examined the relationship between land use and accidents including pedestrian accidents in the City and County of Honolulu. By linking police crash data (1986-1995) with land used data at the parcel level, each accident point was assigned to a land use code including Residential, Visitor Lodging or Other, Education, Religious & Social Institutions, Recreational and Cultural Activities, Military,
Public Services, Manufacturing & Industrial, Commercial & Services, Utilities and Communication, Agriculture, and Vacant/Open land.

There were several findings in the research. First, Visitor Lodging and Commercial land use dominated in pedestrian accidents. Second, the number of pedestrian crashes produced per acre of residential land use was much lower than other land uses.

It was pointed out that the difficult nature of understanding the relationships between land use and traffic accidents. First, accidents occur on roadways, generally not on the adjacent properties. Second, the classification of land use may not be the most appropriate categories for accident analysis. Third, the land use maps do not completely portray all uses of land. For instance, there might be residential uses occurring within commercial or office districts.

In the aforementioned research by Graham (2003a), land use data were not observed at the ward level, so it was proposed that the amount of employment and the amount of population per network node were indicative overall of the relative densities of economic and residential uses.

This research indicated that land use factors should be included in the pedestrian casualty model, in particular, residential land use and commercial land use factors.

Graham’s another study (2003b) showed that children were more likely to encounter traffic accidents where residential populations were high, and that adults were more likely to be hit by cars in large employment centers.

Noland’s aforementioned research (2003) also considered the land use effect on traffic accidents. Dummy variables for land use factors were created based upon the
distribution of observed employment density and population density. Dummy variables for the level of urbanization were also included as a land use variable.

2.3. Roadway Factors

Many researchers have insisted that roadways have an effect on the number of accidents. However, most of the researches used road segments as the dependent variable. The focus of this research is the characteristics of the accident-prone areas, not the roadways. Therefore, only the research related to zonal analysis are mentioned in this section.

Levine’s research (1995b) included roadway factors as a contributing factor. The results showed that three of the roadway variables (the existence of a freeway link, miles of major arterial, and miles of freeway ramp/freeway access) produced positive and significant regression coefficients.

Harruff’s pedestrian fatality analysis (1998) showed there was little correlation of the severity of injuries with the types of roadways.

Graham’s research (2003a) showed that the length of main streets in a ward had a positive effect on the number of pedestrian casualties, but minor roads had a negative effect. His follow-up research (2003b) found that roadway characteristics had little effect on fatality accidents. Road length was, however, associated with increases in serious injuries. Also, the number of junctions and roundabouts in wards were not associated with any casualty types, except slight injuries.

Noland’s aforementioned research (2004) showed that road infrastructure, expressed as the amount of each functional road class within a region, could affect the
casualties. More minor roads in a region were associated with fewer casualties. He suggested that if minor roads tend to have lower speeds, limiting speeds might be effective at reducing casualties.

2.4. Traffic Factors

Some researchers have suggested that traffic factors, in particular traffic volume, are associated with the number of pedestrian accidents. However, traffic volume in a particular area is more difficult to obtain than traffic volume of the roadways. Also, traffic volume data were not available in some of the research, and some of the studies applied proxy variables to represent traffic volume in study areas.

Levine’s aforementioned research (1995b) suggested traffic volume data in the zones would have improved his model, which was not available for his research.

In Petch’s aforementioned research on child safety (2000), the model predicted a positive correlation between average traffic volume and the child pedestrian/cyclist casualty rate. In the case of main streets, a positive association was found between average traffic volume and child pedestrian/cyclist casualty rate.

LaScala (2000)’s aforementioned research also showed that pedestrian injury rates were positively related to high traffic flow.

To estimate traffic flows in each ward, the aforementioned research by Graham (2003a) constructed proxy variables for traffic flow. The level of employment and resident population was as a parameter of traffic volume in the ward and included in the model. The logic was that the relative location of people, jobs, and distances from other wards can provide proxy variables for traffic flow.
The results were mixed. As the employment parameter increased, the incidence of pedestrian casualties increased in most cases. However, in highly developed areas, the incidence fell, probably due to congestion effect on traffic speeds or traffic calming measures. In contrast, proximate population had an increasing effect on pedestrian casualties.

Noland’s aforementioned research (2003) adopted the same method as Graham’s to represent traffic volume. He concluded traffic proxy variables were positively related to the number of accidents in the wards.

2.5. Statistical Properties of Models

Regression analysis has been used for most of the existing studies to relate accidents to a set of independent variables. Regression analysis is used for this research because regression model can estimate the statistical significance of multiple factors (variables) in one model. Each of the factors can be tested its effect on the dependent variable as an additional effect. If one of the variables showed a statistically significant value, it can be interpreted that the variable affects the dependent variable in addition to the rest of the independent variables.

Statistical modeling of accident frequency of a given area can be obtained by multiple linear regression, Poisson regression or negative binomial regression. In this section, the characteristics of these three regression models are reviewed.

2.5.1. Multiple Linear Regression Model

Multiple linear regression models have the following general form:
\[ \lambda = \beta x + \varepsilon \]

Where

\( \lambda \) = Expected mean number of events.

\( x \) = Vectors representing the independent variables.

\( \beta \) = Vectors representing parameters to be estimated.

\( \varepsilon \) = Error terms assumed to be distributed as normal (Mostafa, 1998)

To identify the optimum model, F-value, R-square and mean square error should be used. Individual parameters in the \( \beta \) vector are tested to investigate the null hypothesis that a given parameter is zero using t-statistics. However, multiple linear regression should be used with caution because accident frequency data are non-negative, non-normally distributed and have error terms with unequal variance (Mostafa, 1998). Ng and Abdel-Aty (2000) stated that the multiple regression model was unsuitable for determining the number of accident events.

2.5.2. Poisson Regression Model

Poisson regression is based on the assumption that the dependent variable is Poisson-distributed. Poisson regression models the probability of discrete events such as traffic accidents according to the Poisson process as follows:

\[
\text{Prob}(n_i) = \frac{\lambda_i^{n_i} \exp(-\lambda_i)}{n_i!}
\]
And \( \lambda_i = \exp(\beta x_i) \)

Where,

\( n_i \) is the target number of events on section \( i \) over a period of time \( t \);
\( \lambda_i \) is expected mean number of events;
\( x \) is a vector representing the independent variables of section \( i \); and
\( \beta \) is a vector representing parameters to be estimated;

In Poisson regression, the coefficient vector \( \beta \) can be estimated by a standard maximum likelihood method with the likelihood function, \( L(\beta) \), being

\[
L(\beta) = \prod_{i=1}^{N} \frac{\exp[-\exp(\beta x_i)] \exp(\beta x_i)^{n_i}}{n_i!}
\]

Where, \( N \) is the total number of analysis units (Mostafa, 1998).

2.5.3. Negative Binomial Regression Model

The negative binomial model arises from the Poisson model by specifying:

\[
\lambda = \exp(\beta x_i + \varepsilon_i)
\]

Where,
\( \lambda_i \) is expected mean number of events on section \( i \);

\( \beta \) is a vector representing parameters to be estimated;

\( x_i \) is a vector representing the an independent variable on section \( i \);

\( \varepsilon_i \) is error term, where \( \exp(\varepsilon_i) \) has a gamma distribution with mean 1 and variance \( \alpha^2 \).

The resulting probability distribution is as follows:

\[
\text{Pr}(n_i|\varepsilon_i) = \frac{\exp[-\lambda_i \exp(\varepsilon_i)][\lambda_i \exp(\varepsilon_i)]^n}{n_i!}
\]

Integrating \( \varepsilon \) out of the expression produces the unconditional distribution of \( n \).

The formulation of this distribution is:

\[
\text{Pr}(n_i) = \frac{\Gamma(\theta + n_i)}{(\Gamma(\theta)n_i!)} u_i^\theta (1 - u_i)^\theta
\]

Where,

\[ u_i = \frac{\theta}{\theta + \lambda_i} \] and \[ \theta = \frac{\theta}{\alpha} \]

The corresponding likelihood function is:
$L(\lambda_i) = \prod_{i=1}^{N} \frac{\Gamma(\theta + n_i)}{\Gamma(\theta) n_i!} u_i^\theta (1 - u_i)^{n_i}$

Where $N$ is the total number of sections. This function is maximized to obtain coefficient estimates for $\beta$ and $\alpha$. Compared with the Poisson model, this model has an additional parameter $\alpha$, such that

$$Var[n_i] = E[n_i][1 + \alpha E[n_i]]$$

The choice between the negative binomial model and the Poisson model is determined by the statistical significance of the estimated coefficient $\alpha$. If $\alpha$ is not significantly different from zero (as measured by t-statistics) the negative binomial model simply reduces to a Poisson regression with

$$Var[n_i] = E[n_i]$$

If $\alpha$ is significantly different from zero, the negative binomial is the correct choice and the Poisson model is inappropriate (Mostafa, 1998).

Therefore, in most research on accident analysis, multiple linear regression was not used because of the characteristics of accident data. Poisson regression and negative binomial regression have been used for this type of research. Also, the choice between Poisson regression and negative binomial regression was made based on the value of $\alpha$. 
2.6. Spatial Analysis Technique

Some of the studies used Geographic Information System (GIS) to perform spatial analysis on traffic accidents. GIS is a computer-based system capable of capturing, storing, manipulating, and displaying geographically referenced data, and converting them into spatial information useful in solving complex spatial problems (Yeung, 2002). Using GIS, the relationship between accident occurrence and the characteristics of the areas can be analyzed. Specifically, the following techniques are used in this research.

Density is a function in ArcGIS Spatial Analyst that distributes the quantity or magnitude of point or line observations over a unit of area to create a continuous raster (GIS Dictionary, 2004). Based on accident point data, an accident density map is created by this function.

Zonal statistics is a function in ArcGIS Spatial Analyst that calculates a statistic for each zone of a zone data set based on values from another data set. A single output value is computed for every zone in the input zone data set (GIS Dictionary, 2004). Accident density for each area (census block group) is calculated using Zonal statistics.

Intersect is the topological overlay of two layers that preserves a feature falling within the spatial extent common to both layers (Yeung, 2002). Intersect make it possible to apply the boundaries of census block groups to other datasets such as land use and roadway data.

Spatial Join is a type of table join operations in which fields from one layer’s attribute table are appended to another layer’s attribute table based on the relative
locations of the features in the two layers (GIS Dictionary, 2004). The traffic volume of each area is estimated from AADT point data using this function.

2.7. Summary of The Literature Review

After careful review of the literature, this research found several important aspects that should be considered when performing pedestrian accident analysis.

Low-income areas are positively related to the number of pedestrian accidents. Age distribution in the areas should also be taken into account. The effect of land use types, especially residential areas and commercial areas, needs to be explored. Roadway factors and traffic volume in the areas should be included as factors as well. However, the result of this research might be different from the ones from previous studies. The degree of development, transportation methods, or other geographic conditions can affect the results. The result of the research in the Hillsborough County might not be the same as the result of the other research. For example, some of the previous studies were conducted in England. Railroad transportation is developed in England. Commuters do not necessarily use cars to go to their workplaces. In the Hillsborough County, passenger trail is not available. Most people commute to work by car. Also, the scale of the research is different. Some research used much larger area as the analysis unit than this research. These factors can affect the result of the model.

The multiple linear regression model, which is useful in many situations, is not an appropriate method when using the number of accidents as a dependent variable. Accident data are generally non-negative and non-normally distributed, and they have
error terms with unequal variance. Most studies have used Poisson regression or negative binomial regression to perform accident analysis.

A large body of research exists on the influence of road geometry and demographic factors on traffic accidents including pedestrian accidents. However, fewer studies have examined the effect of land use and traffic volume on the number of accidents, mainly because these data were difficult to obtain. Also, most studies analyzed large geographic area such as ‘ward’ (the average area of each ward is about 14 square kilometers in England). In order to implement effective policy for pedestrians, analysis on smaller geographic areas is important when considering limited funds for pedestrian safety. In addition, zonal analysis assumes that the risk of accident is uniform at all locations within the zone, which is frequently not true. Small scale analysis is necessary to analyze accidents accurately. In this research, a census block group, which is a fairly small geographic area (the average area of census block groups is 4.125 square kilometers in the Hillsborough County) is used as a spatial analysis unit.
CHAPTER 3: METHODOLOGY

3.1. Data Collection

3.1.1. Accident Data

The accident data used in this research was obtained from the Traffic Division at the Hillsborough County for the years 1999 to 2001. The data consist of the accident records reported by the City of Tampa and the Hillsborough County Sheriff’s office. The data included 1,648 pedestrian accident records (Figure 2). They are stored in shapefile format (shapefile is the file format for graphical data used by ArcGIS). Because Hillsborough County has several authorities which report the accidents in different formats, it was difficult to collect all of the pedestrian accident data. This was the largest sample dataset available for the research.
3.1.2. Roadway Data

Roadway data were obtained from the Florida Geographic Data Library (Florida Geographic Data Library, 2004). The roadway data were created in 2001 by Florida Department of Transportation and the Transportation Statistics Office. The data cover all of the roadways owned by the State of Florida. The data are shown in Figure 3. In addition, roadway data which include local roadways were obtained from the Traffic
Division at Hillsborough County. This dataset was created in 2001, and it was stored in shapefile format as well.

Figure 3 State Roadways

3.1.3. Census Data

Census block groups were used as spatial units for the model. As described in Chapter 2, accident analysis on small geographic area is lacking from previous research
efforts. The smallest census unit is census block, but demographic data were not available at census block level. Also, it was expected that most of the census blocks would have no accident value (Even at census block group level, 59 out of 732 block groups had a value of zero). Accordingly, census block groups were used as analysis unit.

In the year of 2000, there were 795 census block groups in the county; and the average area of one census block group was 4.125km². Also, the average number of residents in one census block group was 1,256. The census data were created by the U.S. Census Bureau, and stored in shapefile format by the Geoplan Center at the University of Florida (Geoplan Center, 2004). The census data used in this research include average household income, population density, and population by age groups in each block group, which are used as demographic variables in the model. The boundaries of census blocks are shown in Figure 4.
3.1.4. Traffic Volume Data

Traffic volume data were provided by the Metropolitan Planning Organization. The data include Annual Average Daily Traffic (AADT) measured at 551 stations from 1999 to 2001 in the Hillsborough County (MPO, personal communication). This value is used as a parameter of the traffic volume for each census block. The procedures aggregating AADT into each block groups is described in a later section.
3.1.5. Land Use Data

Land use data were collected to examine land use distribution in each census block group. The land use data were obtained from South West Florida Water Management District (SWFWMD, 2004). Land use data were created in 1999, and stored in shapefile format. All of the areas in Hillsborough County were categorized by more than 40 land use types based on Florida Department of Transportation Florida Land Use and Land Cover Classification System (FLUCCS). Land use map is shown in Figure 5. Zoomed area shows the neighborhood of the University of South Florida.

Among those are ‘Commercial and Service’, ‘Residential High Density’, ‘Residential Middle Density’, and ‘Residential Low Density’. Because one of the hypotheses is that these concentrated residential and commercial land use affect the number of pedestrian accidents, these four land use types are used as independent variables in the model.
3.2. Data Processing

3.2.1. Software Used

The GIS software used was ArcGIS 8.3, developed by ESRI. The procedure of combining the accident points, census, land use, roadway, and traffic volume data is described in the following section. As for statistical analysis, LIMDEP was used in this
research. LIMDEP was developed by Econometric Software Corporation. Regression models in this research were examined using this software.

3.2.2. Exclusion of The Areas

Some of the census block groups were removed from analysis. There were eight census block groups with no residents. These zones were excluded because demographic analysis is not possible for those block groups. Also, 53 block groups within the jurisdiction of Plant City and Temple Terrace were removed as well. This was because although these authorities investigate traffic accidents, their reports were not available for this research. In order to avoid bias in the data, these census blocks should not be used for analysis. As a result, 732 out of 795 block groups were used for analysis. Excluded areas are shown in Figure 6.
3.2.3. Create Accident Density Map

Accident points have to be counted for each census block group. As seen Figure 2 and Figure 4, Accident data are points, and census block group data are areas in nature. The easiest way to count accidents is to overlay these two datasets and count the number of accident points for each census block group.

However, there was a problem with the accident point data. All of the accident points were geocoded based on intersections. Even though some of the accidents

Figure 6 Excluded Census Block Groups

31
occurred somewhere between the intersections, they were geocoded on the closest intersection. In other words, coded accident points do not represent the exact location where the accidents occurred.

Also, another problem was identified which is related to the characteristics of point data. All of the accident points are geocoded on only one of the census block groups, in spite of the fact that some of the accidents occur at the boundary of census block groups. These accidents are counted for either one of the census block groups. However, accidents occur at the boundaries of census block groups, not at one of the census block groups. In order to take account of adjacent areas, accident points have to be spread out.

This research decided to create an accident density map to deal with these issues. All of the areas in Hillsborough County were divided into 30m by 30m cells. For each 30m by 30m cell, the number of accidents within a 500m radius was counted. The search radius was set at 500 meters because a previous study by Abdalla (1997) showed most of the pedestrian accidents occurred within 500 meters of the casualty’s home.

There are two options in ArcGIS 8.3 to calculate the density of the cells, which are simple method and Kernel method.

For each cell, the simple method adds up the number of accidents which occurred within a circle of 500 meters radius of the cell. The distance from the cell is not considered for this method. For example, an accident which occurred at the distance of 499 meters (from the center of the cell) is counted as same weight as an accident at the distance of five meters. This method is not appropriate to represent accident point density.
A cell value can be very high though accident does not occur nearby the cell. Also, a cell value can be low even if an accident point is located exactly at the cell.

On the other hand, the Kernel method calculates cell values using the distance of accident points from the center of the cell. The Kernel method creates a density map with a smooth surface in which the density at each location reflects the concentration of points in the surrounding area. Distances are measured from the center of the cell to each observation that falls within 500 meters. Each accident point contributes to the density value of the cell based on its distance from the center. Nearby accidents are given more weight in the density calculation than those further away. For example, an accident which occurred at five meters from the center of the cell is counted with much more weight than an accident at 499 meters. Therefore, the Kernel method was used to create the accident density map. The resulting map is shown in Figure 7.
3.2.4. Calculate Accident Value for Census Block Groups

The average value of all of the cells in each census block group was calculated using zonal statistics. The output value is the number of accidents per km² for the census block groups. This value ranged from 0 to 35.72 accidents per km². It should be mentioned that 391 of the 732 block groups had a value of less than 1, and 59 out of 732 block groups had a value of zero. These low values have to be considered when selecting
an appropriate statistical model. This is discussed in the ‘statistical analysis’ section of this chapter.

3.2.5. Development of Land Use Variables

As mentioned before, this research uses four types of independent variables contributing to pedestrian accidents. Land use variables were developed as follows. Census block groups were intersected with land use data. Figure 8 is a map which represents the four land use types used as independent variables. Zoomed area shows the neighborhood of the University of South Florida.
Next, for each block group, the total area was calculated as well as the total areas of ‘Commercial and Service’, ‘Residential High Density’, ‘Residential Medium Density’, and ‘Residential Low Density’. The area of ‘Commercial and Service’ was divided by the total area of the census block group, and multiplied by 100, which represents the percentage of commercial and service land use in a census block group. This value formed one independent variable in the model.

As for residential land use, the area was calculated based on resident density. It is a reasonable assumption that there are more pedestrians in the neighborhoods of high density residential areas than ones in low density areas. If there are more pedestrians on the street, those areas are more likely to have higher accident rates. This should be taken into account.

According to the definition of the land use data, ‘Residential High Density’ areas are composed of areas with six or more dwelling units per acre. ‘Residential Medium Density’ areas have two to five dwelling units per acre. Also, ‘Residential Low Density’ areas are less than two dwelling units per acre (FDOT, 2004). Using this information, the residential land use variable was calculated as follows:

\[
\frac{\text{The area of ‘Residential High Density’} \times 3 + \text{the area of ‘Residential Med Density’} \times 2 + \text{the area of ‘Residential Low Density’} \times 1}{\text{the total area of the block group}} \times 100
\]
This value was included in the model as an independent variable.

3.2.6. Development of Roadway Variable

In preliminary research, all of the roadways including local roads were calculated as roadway variable. However, it was found that the length of all roadways was highly correlated with the residential land use variable. It was also found that the length of state roadways was less correlated with any of the independent variables. In addition, preliminary research found that 68.3 % of the accidents (1126 out of 1648 accidents) occurred along state roadways. Therefore, it was decided to use the length of state roadways as roadway variable.

The roadway shapefile, which included every state roadway, was intersected with the census block group shapefile.

The total length of the state roadways in each block group was calculated. Next, this value was divided by the total land area of each block group to obtain roadway density in meters per km². It is obvious that a census block group with a large area is more likely to have long roadways. It does not mean, however, that the roadways are concentrated in those large block groups. Therefore, in order to explore the concentration of the roadways, the length of roadway was divided by land area. This value was used as a roadway variable in the model.

3.2.7. Development of Traffic Variable
Using traffic count data from the Metropolitan Planning Organization, traffic count values were assigned to census block groups. Figure 9 shows the locations of the points where AADT was measured.

Spatial join was used to assign AADT values to each census block group. If a census block group does not have any AADT point inside the area, the value of the closest AADT point was assigned to each census block group. The average distance from a census block group to AADT point was 355 meters. If a census block group has AADT
point inside the area, the value of AADT was applied to the census block group. For the census block groups with multiple AADT points inside, the average of the AADT values were calculated and used as AADT values for the census block groups. It is assumed that these values represent the traffic volume of the census block group.

3.2.8. Development of Demographic Variables

Each census block group has demographic data based on the 2000 census. Population, average household income, and population by age were used in the research. First, average household income was included in the model. Second, population density, children-rate, and elderly-rate were calculated as follows:

Population density = total population / total area (km²)
Children-ratio = population at the age of 0 ~ 17 / total population
Elderly-ratio = population of more than 65 years old / total population

These values were included in the model as well.

3.3. Statistical Analysis

3.3.1. Distribution Fitting

The dependent variable in the model is the average number of pedestrian accidents from 1999 to 2001 in each census block group. Prior to the statistical modeling, the general shape of the average number of pedestrian accidents was explored in order to
provide the basis for understanding distribution of the data. Figure 10 shows the histogram for average number of pedestrian accidents in census block groups. As mentioned before, the search radius of the density map was set at 500m. In addition to this, two search radii (250m and 1000m) were tested to see whether the search radius changes the distribution of the dependent variable. For reference, two additional histograms are shown in Figure 11 (search radius of 250m) and Figure 12 (search radius of 1000m).

Census Block Groups

| Census Block Groups | n = 732 | Mean = 2.78 | Max = 35.72 | Min = 0 | Standard Deviation = 4.07 |

Figure 10 Distribution of The Dependent Variable
Census Block               n = 732  Mean = 2.75  Max = 40.96  Min = 0  Standard Deviation = 4.58
Groups

Figure 11 Distribution with A Radius of 250m
In Figure 10, it is apparent that a large number of census block groups have no or very low values of pedestrian accidents. The distribution seemed to follow the Poisson distribution. Also, as seen in Figure 11 and Figure 12, the distribution was similar to the one in Figure 10.

3.3.2. Selection of Statistical Model
As previous researchers suggested, in general, accident frequency is non-negative, non-normally distributed and it has error terms with unequal variance. These are violations of the statistical assumptions of the multiple linear regression models such as a normal distribution and homoscedastic assumption on error terms. Violation of the assumptions can invalidate the hypothesis tests concerning the significance of the parameters (Jovanis & Chang, 1986). In addition, the multiple linear regression model might predict a negative number of accidents that never occurs in reality. Preliminary research using multiple linear regression showed that 94 out of 732 census block groups predicted negative values. Therefore, other types of models should be considered.

Due to the problems associated with the multiple linear regression model, the Poisson regression model should be considered. The Poisson regression possesses most of the desirable statistical properties in describing accident frequency (Jovanis et al., 1986, Miau, et al, 1993, Shanker, et al, 1995). However, Poisson regression requires the variance of the data to be equal to the mean. When the mean and the variance of the data are not approximately equal, the model coefficients are biased (Peng, 2004). When the variance is larger than the mean, it is termed overdispersion. This discrepancy can be overcome by an additional assumption on the expected mean, which is a gamma distribution assumption. It is assumed that the number of accident events has a Poisson distribution with the expected mean of a gamma distribution. As such, the number of accident events can be regarded as having a negative binomial distribution (Cox, 1983; Land et al., 1996; McCullagh et al, 1989). The negative binomial regression model is particularly useful in accounting for the overdispersion (Land et al., 1996).
The mean and variance of the data were calculated for the dependent variable (number of accidents / km²). for this research. The mean and variance were 4.072 and 16,579, respectively. The variance is much larger than the mean, which suggests that the results of the modeling would be biased if this research used the Poisson regression model.

Also, the choice between the Negative Binomial model and the Poisson model can largely be determined by the statistical significance of the estimated coefficient $\alpha$. If $\alpha$ is not significantly different from zero (as measured by t-statistics) the negative binomial model simply reduces to a Poisson regression. If $\alpha$ is significantly different from zero, the negative binomial is the correct choice and the Poisson model is inappropriate (Mostafa, 1998). $\alpha$ value of the data, which is critical to select an appropriate model, is assessed in the next chapter.
CHAPTER 4: RESULTS AND DISCUSSION

4.1. Statistical Properties of The Variables

Before running regression models, statistical properties of each variable were reviewed, which is shown in Table 1. Mean, standard deviation, minimum value, maximal values were examined.
Table 1 Statistical Properties of Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>min.</th>
<th>max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accidents (number / km²)</td>
<td>2.781792</td>
<td>4.071728</td>
<td>0</td>
<td>35.72015</td>
</tr>
<tr>
<td>Children rate ratio</td>
<td>0.244663</td>
<td>0.089762</td>
<td>0</td>
<td>0.617</td>
</tr>
<tr>
<td>Elderly ratio</td>
<td>0.130598</td>
<td>0.123647</td>
<td>0</td>
<td>0.899</td>
</tr>
<tr>
<td>Average income (dollars)</td>
<td>50120.61</td>
<td>24821.19</td>
<td>0</td>
<td>200001</td>
</tr>
<tr>
<td>Commercial (%)</td>
<td>10.53521</td>
<td>14.23568</td>
<td>0</td>
<td>88.80763</td>
</tr>
<tr>
<td>Residential land use (%)</td>
<td>142.7811</td>
<td>89.52254</td>
<td>0</td>
<td>300</td>
</tr>
<tr>
<td>Road density (m / km²)</td>
<td>1693.212</td>
<td>1667.707</td>
<td>0</td>
<td>13820.22</td>
</tr>
<tr>
<td>Traffic volume (vehicles / day)</td>
<td>31313.22</td>
<td>27996.13</td>
<td>1091</td>
<td>163000</td>
</tr>
</tbody>
</table>

4.2. Result of Negative Binomial Regression Model

Table 2 represents the accident model obtained by negative binomial regression. The dispersion parameter, $\alpha$, is highly significant (as t-value is 19.909 which is much greater than 1.96) at the 95% confidence interval. Therefore, the mean varies from the variance significantly, and this dataset is overdispersed. This value confirms the
appropriateness of the negative binomial regression relative to the Poisson regression for modeling this dataset.

In order to measure the overall goodness-of-fit, the deviance value \(2(\text{LL}(\beta)-(0))\) which follows the Chi-square distribution has been used as suggested by Agresti (1990). This Chi-square test of the deviance value \(310646.6\) at degree of freedom at 6, which is analogues to the F-test in linear regression modeling, strongly rejects the null hypothesis that the obtained model has explanatory power equal to that of the model with the constant term only. This value therefore shows a good overall statistical fit.

Also, the coefficient of correlations was estimated for each independent variable. The result is shown in Table 3.
<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Value</th>
<th>Standard Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.9646</td>
<td>0.2293</td>
<td>17.292</td>
<td>0.00</td>
</tr>
<tr>
<td>Children resident (0 to 17 years) ratio</td>
<td>0.7890E-01</td>
<td>0.5385</td>
<td>0.147</td>
<td>0.88</td>
</tr>
<tr>
<td>Elderly resident (65 years or more) ratio</td>
<td>-0.9537</td>
<td>0.4114</td>
<td>-2.318</td>
<td>0.02</td>
</tr>
<tr>
<td>Average household income</td>
<td>-0.1838E-04</td>
<td>0.1665E-05</td>
<td>-11.041</td>
<td>0.00</td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.4476E-01</td>
<td>0.4235E-02</td>
<td>10.570</td>
<td>0.00</td>
</tr>
<tr>
<td>Residential land use</td>
<td>0.8409E-02</td>
<td>0.5316E-03</td>
<td>15.819</td>
<td>0.00</td>
</tr>
<tr>
<td>Road density</td>
<td>0.1598E-03</td>
<td>0.3250E-04</td>
<td>4.916</td>
<td>0.00</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>3.7834E-06</td>
<td>0.1739E-05</td>
<td>2.212</td>
<td>0.03</td>
</tr>
<tr>
<td>Chi squared</td>
<td></td>
<td></td>
<td></td>
<td>310646.6</td>
</tr>
<tr>
<td>α</td>
<td></td>
<td></td>
<td></td>
<td>19.909</td>
</tr>
</tbody>
</table>
Table 3 Coefficient of Correlations for Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Accidents</th>
<th>Children ratio</th>
<th>Elderly ratio</th>
<th>Medium family income</th>
<th>Commercial and Service land use</th>
<th>Residential land use</th>
<th>Road density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children rate</td>
<td>0.065</td>
<td>-0.083</td>
<td>-0.509</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elderly rate</td>
<td>0.002</td>
<td>-0.055</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average household income</td>
<td>-0.405</td>
<td>-0.055</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.498</td>
<td>-0.131</td>
<td>0.026</td>
<td>-0.328</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential land use</td>
<td>0.204</td>
<td>0.013</td>
<td>0.076</td>
<td>0.037</td>
<td>-0.104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road density</td>
<td>0.475</td>
<td>-0.15</td>
<td>0.04</td>
<td>-0.155</td>
<td>0.544</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Traffic volume</td>
<td>0.230</td>
<td>-0.113</td>
<td>-0.099</td>
<td>-0.063</td>
<td>0.185</td>
<td>-0.012</td>
<td>0.119</td>
</tr>
</tbody>
</table>

4.3. Discussion of Each Variable

4.3.1. Demographic Variables

Average household income, children resident ratio, and elderly resident ratio were included in the model as demographic variables. Table 2 showed that average household income has a strong negative effect on the number of pedestrian accidents in census block groups. The t-value of -11.041 is much lower than -1.96, which is the 95% confidence level of significance. This value suggests that an increase in the average household income lowers pedestrian accidents in a census block group.
The elderly resident rate showed a t-value of -2.318. This value suggests a negative relationship between the elderly resident rate and pedestrian accidents. This result was against the hypothesis that an area with high elderly resident ratio experiences more accidents than others. A possible interpretation could be that elderly people spend more time at home than the other population and they are less likely to encounter accidents. It could also be suggested that elderly people are more careful crossing the streets.

Harruff’s research (1998) showed the oldest age group has the highest pedestrian fatality rates, which differs from the result of this research. A probable explanation is that this research used all pedestrian accidents regardless of the severity of the injuries. If fatality accidents were used as the dependent variable, the result would have been similar to the one in Harruff’s research, because elderly people are likely to die of injuries which younger people might survive.

The children resident ratio did not indicate a statistically significant result. The children resident ratio was expected to have a positive relationship with pedestrian accidents based on the assumption that children are generally more careless than adults. It could be that children at a very young age are accompanied by their parents, and the parents watch out for cars. The result could have been different if the research had used children over 10 years old. Another factor could be that some of the children go to school using a school bus, and they do not walk around the neighborhood very often.

In Petch’s research (2000), the model predicted a positive correlation between average traffic volume and child pedestrian/cyclist casualty rate. Only children casualties
were used as the dependent variable by Petch’s research. It could interpreted that children might be more vulnerable in congested areas than the other age groups.

However, there was a correlation between the children resident ratio and the elderly resident ratio. Table 3 indicated the coefficient of correlation was -0.509. This may result in multicollinearity. Multicollinearity exists when two or more of the independent variables used in regression are moderately or highly correlated. High correlations among independent variables increase the likelihood of rounding errors in the calculations of the β estimates and standard errors. Multicollinearity can also have an effect on the signs of the parameter estimates (Mendenhall et al, 2003).

In order to avoid multicollinearity and improve the accuracy of the model, this research assessed two additional accident models. The first model used the same variables as the ones in Table 2 except the children resident ratio variable, because the t-value of the children ratio (0.147) was not statistically significant. The result is shown in Table 4. The second model used the children resident ratio and the elderly resident ratio as one variable. The value of the elderly resident ratio in each census block was simply added to the value of the children resident ratio, and a new valuable, vulnerable population ratio was created. This research hypothesized that both elderly people and children were likely to encounter pedestrian accidents, and this model treated these two age groups together. The result is shown in Table 5.
Table 4 Regression Result without Children Resident Rate

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Value</th>
<th>Standard Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.9900</td>
<td>0.1503</td>
<td>26.540</td>
<td>0.00</td>
</tr>
<tr>
<td>Elderly residents (65 years or more) ratio</td>
<td>-0.9799</td>
<td>0.3707</td>
<td>-2.644</td>
<td>-0.01</td>
</tr>
<tr>
<td>Average household income</td>
<td>-1.8845E-04</td>
<td>0.1605E-05</td>
<td>-11.497</td>
<td>0.00</td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.4470E-01</td>
<td>0.4211E-02</td>
<td>10.613</td>
<td>0.00</td>
</tr>
<tr>
<td>Residential land use</td>
<td>0.8418E-02</td>
<td>0.5284E-03</td>
<td>15.931</td>
<td>0.00</td>
</tr>
<tr>
<td>Road density</td>
<td>0.1597E-03</td>
<td>0.3251E-04</td>
<td>4.913</td>
<td>0.00</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>0.3849E-05</td>
<td>0.1738E-05</td>
<td>2.215</td>
<td>0.02</td>
</tr>
<tr>
<td>Chi squared</td>
<td></td>
<td></td>
<td>310646.6</td>
<td></td>
</tr>
</tbody>
</table>
Table 5 Regression Result with Vulnerable Resident Rate

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Value</th>
<th>Standard Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.1395</td>
<td>0.2155</td>
<td>19.213</td>
<td>0.00</td>
</tr>
<tr>
<td>Vulnerable population ratio (0~17 years old + 65 years or more)</td>
<td>-0.6300</td>
<td>0.4101</td>
<td>-1.536</td>
<td>0.12</td>
</tr>
<tr>
<td>Average household income</td>
<td>-0.1891E-04</td>
<td>0.1636E-05</td>
<td>-11.561</td>
<td>0.00</td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.4388E-01</td>
<td>0.4218E-02</td>
<td>10.403</td>
<td>0.00</td>
</tr>
<tr>
<td>Residential land use</td>
<td>0.8422E-02</td>
<td>0.5352E-03</td>
<td>15.735</td>
<td>0.00</td>
</tr>
<tr>
<td>Road density</td>
<td>0.1547E-03</td>
<td>0.3244E-04</td>
<td>4.767</td>
<td>0.00</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>0.3912E-05</td>
<td>0.1725E-05</td>
<td>2.268</td>
<td>0.02</td>
</tr>
<tr>
<td>Chi squared</td>
<td></td>
<td></td>
<td></td>
<td>310642.8</td>
</tr>
</tbody>
</table>

Table 4 showed the same value of Chi-squared test of deviance value, 310646.6. This value indicated that children resident rate did not improve the model. Therefore, this variable should not be included.

The vulnerable population rate, which was shown in Table 5 did not show a statistically significant result. The t value of the vulnerable population rate was -1.536, which is higher than critical value of 95% confidence level, -1.96. Therefore, only elderly resident ratio should be used as an age factor.

It should be mentioned that there are a significant number of travelers or temporarily residents in the State of Florida. They were not counted as residents in census block groups, but they could also encounter accidents on the street as well as the residents.
The result might have been different if this research had included travelers or temporarily citizens.

4.3.2. Land Use Variables

Residential land use and commercial and service land use were included as land use variables. In Table 4, residential land use showed a t-value of 15.931. This result indicates that residential land use had an increasing effect on the number of pedestrian accidents. As described in Chapter 3, this research adjusted the area of residential land use based on the density of the residential land use type. High density residential areas were multiplied by 3, and middle density residential areas were multiplied by 2, while low density areas were left intact.

Because of this adjustment, this variable represented the density of residents in census block groups, as well as the percentage of residential areas. Stated differently, we could attain a similar result using population density instead of adjusted residential land use. We also might get a different result simply by adopting the percentage of residential land use without adjustment. Two additional regression models were developed using population density and the percentage of residential land use, respectively.
Table 6 Negative Binomial Model with Population Density

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Value</th>
<th>Standard Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>4.1971</td>
<td>0.1649</td>
<td>25.455</td>
<td>0.00</td>
</tr>
<tr>
<td>Elderly residents (65 years or more) ratio</td>
<td>0.2722</td>
<td>0.4441</td>
<td>0.613</td>
<td>0.54</td>
</tr>
<tr>
<td>Average income</td>
<td>-0.1525E-04</td>
<td>0.1732E-05</td>
<td>-8.802</td>
<td>0.00</td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.3609E-01</td>
<td>0.4362E-02</td>
<td>8.275</td>
<td>0.00</td>
</tr>
<tr>
<td>Population density</td>
<td>0.5613E-03</td>
<td>0.4962E-04</td>
<td>11.313</td>
<td>0.00</td>
</tr>
<tr>
<td>Road density</td>
<td>0.1835E-03</td>
<td>0.3329E-04</td>
<td>5.512</td>
<td>0.00</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>0.5613E-03</td>
<td>0.4962E-04</td>
<td>1.707</td>
<td>0.09</td>
</tr>
<tr>
<td>Chi squared</td>
<td></td>
<td></td>
<td>310592.6</td>
<td></td>
</tr>
</tbody>
</table>
Table 7 Negative Binomial Model with Residential Land Use Percentage

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Value</th>
<th>Standard Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.8193</td>
<td>0.1614</td>
<td>23.656</td>
<td>0.00</td>
</tr>
<tr>
<td>Elderly residents (65 years or more) ratio</td>
<td>-0.9005</td>
<td>0.3753</td>
<td>-2.400</td>
<td>0.02</td>
</tr>
<tr>
<td>Average income</td>
<td>-0.1879E-04</td>
<td>0.1586E-05</td>
<td>-11.849</td>
<td>0.00</td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.4640E-01</td>
<td>0.4360E-02</td>
<td>10.644</td>
<td>0.00</td>
</tr>
<tr>
<td>Residential land use (not adjusted)</td>
<td>0.2487E-01</td>
<td>0.1654E-02</td>
<td>15.034</td>
<td>0.00</td>
</tr>
<tr>
<td>Road density</td>
<td>0.1831E-03</td>
<td>0.3360E-04</td>
<td>5.538</td>
<td>0.00</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>0.4554E-05</td>
<td>0.1793E-05</td>
<td>2.540</td>
<td>0.01</td>
</tr>
<tr>
<td>Chi squared</td>
<td></td>
<td></td>
<td>310638.6</td>
<td></td>
</tr>
</tbody>
</table>

As expected, Table 6 showed a similar result to the one with adjusted residential land use (Table 4). However, elderly resident ratio and traffic volume did not show significant results at the 95% confidence level. Absolute values decreased for all of the variables, and the value of Chi-squared fell from 310646.6 to 310592.6. Population density showed a t value of 11.313, which suggests an increase in population density is positively related to the number of pedestrian accidents. The overall accuracy of the model was slightly decreased by replacing adjusted land use by population density.

Table 7 showed a very similar result to the one with adjusted residential land use (Table 4). All of the t values showed the same sign, and each of them was very close, though the value of Chi squared test decreased slightly (310646.6 to 310638.6). This can
be interpreted that pedestrians tend to encounter accidents nearby residential areas, and accident prone areas are not necessarily high density areas. It was found that the adjustment of land use percentage did not improve the model.

Levine (1995a) showed that most motor vehicle accidents (not only pedestrian accidents) were closer to employment centers than to residential areas, which do not agree with the result of this research. This research showed both of residential land use and commercial land use increase the number of accidents. This difference could be explained that car-on-car accidents and pedestrian accidents have a different spatial distribution.

Graham’s study (2003b) showed that children were more likely to encounter pedestrian accidents where residential population are high, and adults were more likely to be hit in large employment centers. Land use effect might vary depending on the age groups.

In addition, the result could have been different if different coefficients had been used to represent resident density. For instance, ‘Residential High Density’ areas were multiplied by three. According to the definition of the land use data, these areas have six or more dwelling units per acre. Some of the multi-story apartments have much more than six dwelling units per acre. In that case, high density areas should be multiplied by a higher coefficient than 3, because more pedestrians are expected in these areas. The information on the details of land use types was not available for the research. Such information will be helpful for future research.
4.3.3. Roadway Variable

In Table 4, the t-value of the roadway variable was 4.913. This value suggested that state roadway density is positively correlated to the number of pedestrian accidents. It was anticipated because preliminary research showed that 68.3% of accidents occurred along the state roadways.

Table 3 showed that commercial and service land use had some correlation with the roadway variable. The coefficient of correlation was 0.544. This value was the highest of all the correlations. Again, there was a concern that these two variables had multicollinearity. This correlation could be explained by the fact that, in general, commercial sites are developed along major roadways. This research used only state roadways, most of which are main streets. Consequently, this correlation might be unavoidable.

As described in Chapter 3, all types of roadways were used as roadway variable in preliminary research. However, it was found that the density of the all roadways was strongly correlated with the residential land use variable. The coefficient of correlation was approximately 0.70, which was higher than the value between state roadways and commercial and service land use. This correlation might be due to the fact that residential areas generally have a number of narrow and complicated roads inside. It is safe to say that state roadways and commercial and service land use were less correlated compared to all roadways and residential land use. Therefore, the density of state roadways was used in this research.

The effect of the roadway density agrees with the result of previous research efforts, though each research measured the roadway factor in its own method.
This research used the density of state roadways as a roadway variable. However, road geometric and environmental factors are also important to analyze pedestrian accidents. For example, the number of lanes, the median width, sidewalks, or the brightness of the roadways (at night) would have a significant effect on pedestrian accidents. These data were not available at a census block group level. It is recommended that future research addresses these factors.

4.3.4. Traffic Variable

According to Table 4, the traffic variable indicated a t-value of 2.215, which was higher than 1.96 (95% confidence level of significance). This value indicates that traffic volume in census block groups has a positive effect on the number of pedestrian accidents.

As explained in Chapter 3, point data of traffic counts were assigned to each census block group. 289 census block groups had count data inside them, but 443 census block groups did not have count data. For these block groups, the value of traffic count that is closest to the block groups was applied to represent traffic volume. It is uncertain that these values can be used as indicators of traffic volume for the whole block group.

Therefore, additional regression models were tested. This model tested only the block groups that had traffic count data in their area. In this model, 289 census block groups formed a dependent variable, and selected census block groups are shown in Figure 13. The result of negative binomial regression model is shown in Table 8.
Figure 13 Census Block Groups with Traffic Count Data
Table 8 Regression Results for Block Groups with Traffic Count Data inside

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Value</th>
<th>Standard Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.7108</td>
<td>0.2114</td>
<td>17.584</td>
<td>0.00</td>
</tr>
<tr>
<td>Elderly residents (65 years or more) ratio</td>
<td>-0.07646E-01</td>
<td>0.4862</td>
<td>-1.573</td>
<td>0.12</td>
</tr>
<tr>
<td>Average household income</td>
<td>-0.01499E-04</td>
<td>0.2426E-05</td>
<td>-6.178</td>
<td>0.00</td>
</tr>
<tr>
<td>Commercial and service land use</td>
<td>0.4234E-01</td>
<td>0.4928E-02</td>
<td>8.591</td>
<td>0.00</td>
</tr>
<tr>
<td>Residential land use</td>
<td>0.9803E-02</td>
<td>0.8207E-03</td>
<td>11.946</td>
<td>0.00</td>
</tr>
<tr>
<td>Road density</td>
<td>0.1622E-03</td>
<td>0.4482E-04</td>
<td>3.618</td>
<td>0.00</td>
</tr>
<tr>
<td>Traffic volume</td>
<td>0.2777E-05</td>
<td>0.2096E-05</td>
<td>1.325</td>
<td>0.19</td>
</tr>
<tr>
<td>Chi squared</td>
<td></td>
<td></td>
<td>146801.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 8 showed a different result in terms of the traffic variable. Compared to the model including all census block groups (Table 4), the t value dropped from 2.212 to 1.325. This value is smaller than 1.96, a critical value of 95% confidence level. This value does not reject the null hypothesis that traffic variable had no relationship with the number of pedestrian accidents.

This result could be interpreted in several ways. First, a census block group was too small as analysis units. As seen in Figure 13, some of the census block groups had only one point inside their area. It may not be appropriate to measure the traffic volume of the entire zone from the value of a single location. The result would have been different if the analysis units were set at larger areas, such as census tracts, because more
data are obtained per one analysis unit. Lascala (2000) found a positive relationship between traffic flow and pedestrian injury rates at census tract level. However, it was mentioned that the data points for traffic counts were not randomly distributed, because those points included only accident-prone areas and high traffic volume areas.

It could also be suggested that traffic volume actually does not have a strong effect on pedestrian accidents. For example, as Graham (2003a) suggested, high traffic volume areas are not necessarily accident prone areas. In congested streets, vehicle’s speed is generally low, so it is unlikely that pedestrians are involved in accidents. Consequently, census block groups with very high traffic volume might not have many accidents. In addition to this, the traffic regulation such as traffic calming or traffic lights might have affected the result.

In any case, this research could not reach a conclusion about the effect of traffic volume on pedestrian accidents. Understanding the traffic factor will be an important theme for future research.
CHAPTER 5: CONCLUSIONS

5.1. Findings

The objective of the research was to determine the effects of contributing factors on the number of pedestrian accidents in the Hillsborough County. Five hypotheses were proposed to examine these factors. Based on accident records from 1999 to 2001, an accident density map was created, and the value of the accident density was assigned to each census block group. Accident models were developed using negative binomial regression to model the relationship between contributing factors and pedestrian accidents.

The result showed that commercial and service land use, residential land use, and the density of state roadways increases the number of pedestrian accidents. It was also found that average household income and elderly resident ratio lowers the number of pedestrian accidents. Children resident ratio did not show a statistically significant result. Also, population density was used instead of residential land use variable in the additional model, and the result was a similar one. In addition, percentage of residential use showed a similar result to the one with adjusted residential land use by the density of residents.

This research gives useful information on transport and land use planning policies. The study area was in the Hillsborough County, but the findings of the research will be
applicable to the other parts of the country, in particular developing metropolitan areas where public transit is not developed.

5.2. Limitations and Assumptions

It should be mentioned that there were some limitations in the research. First, although the accident data used for the research were the best dataset available, they did not include all of the pedestrian accidents occurred from 1999 to 2001. In addition, the area of the Plant City and the Temple Terrace was removed from analysis because accident records were not available. Second, accident data used to create accident density map were geocoded on intersections, so the location of the accident points was not completely accurate. Third, it was assumed that traffic volume in an area can be measured by the AADT value of the points in the area. The effect of traffic volume was not determined because it is not certain that this method can be used as a parameter for traffic volume for the area. Fourth, this research assumed that pedestrian accidents occur within 500 meters of the casualty’s home.

5.3. Recommendation for Future Research

This research used all kinds of pedestrian accidents regardless of the severity of the accidents. It is recommended that future research address this factor. For example, research on fatality accidents will give useful information.

The density of state roadway was used as a roadway variable. However, road geometric and environmental factors are also important to analyze pedestrian accidents
such as the number of lanes, the median width, sidewalks, or lights. This type of information should be included in future research.

Finally, it is recommended that future research analyze the effect of traffic volume in detail. In this research, some of the census block groups had only one point inside their area. It may not be appropriate to measure the traffic volume of the entire zone from the value of a single location. Using census tracts as analysis units might show a different result. Also, the effect of congestion and traffic regulation should be explored.
REFERENCES


ArcGIS Desktop Help (2004) Retrieved on 08/30/04, from


