Spatial variation in local road pedestrian and bicycle crashes

*__Abram Musinguzi__
Graduate Research Assistant
Department of Civil Engineering
Tennessee State University
3500 John A Merritt Blvd
Nashville, TN 37209
Phone: (615) 943-7811
Email: mailtoabram@gmail.com

Deo Chimba, Ph.D., P.E., PTOE.
Assistant Professor
Department of Civil Engineering
Tennessee State University

2015 Annual ESRI International User Conference
San Diego, California
July 20-24, 2015

Abstract

Local roads represent a large proportion of pedestrians and bicyclists since they provide access to adjacent land. Therefore, identification of high pedestrian and bicycle crash zones on local roads to enhance safety is essential to develop and implement effective countermeasures. Application of Geographic Information Systems (GIS) has gained popularity in traffic safety as they offer immense potential to detect high crash locations. This paper conducted cluster analysis in GIS to identify spatial patterns and high concentration of local road pedestrian crash zones based on statewide crash data collected in Tennessee from 2008 to 2012. Poisson distribution was developed to ascertain whether crashes create clustered pattern by comparing the frequency distribution of observed block groups containing many crashes and those of expected (random) distribution. Chi square analysis was used to test whether two distributions are significantly different. The GIS kernel density tool was applied to create crash density map and locate crash clusters. The findings of this study indicate that pedestrian crash clusters are associated with low-income households, households below poverty level, population with education less than high school and households without vehicles.

1. BACKGROUND

Pedestrian and bicycle safety is a critical issue in the effort to promote walking and bicycling. Yet, statistics indicate that local roads represent the majority of pedestrian and bicycle crashes nationally. In the United States, 42 percent of pedestrian crashes occur on local streets and the largest portion of bicycle crashes (34 percent) occur on local streets [1]. The reason for this is that, exposure levels are likely to be high for the local roads, since their primary role is to provide direct access to land uses, a function that makes them the most numerous type of road in the classification system, and hence most likely to see high numbers of pedestrians and bicyclists.

Identifying high crash zones can help in understanding the factors that front high risks for pedestrians and bicyclists. Several analytical techniques and tools are available to identify high crash zones. However, in recent years, safety research has widely focused on the application of GIS to identify high crash zones. GIS turns statistical data, such as traffic crashes, and geographic data, such as roads and crash locations, into meaningful information for spatial analysis and mapping [2]**.** The identification of traffic accident hot spots provides an insight into the casual factors and is an essential step for appropriate allocation of safety improvements resources [3]. Moreover, identification of high crash zones provides better understanding of spatial patterns and clusters in crash data and this enhances the development of effective safety improvement strategies.

This paper adds to existing literature by introducing a GIS based tool to identify high crash zones for pedestrian crashes occurring on local roads using statewide crash data collected in Tennessee from 2008 to 2012 and demonstrating its application using results from Shelby County. The main goal of this work is to determine whether there is evidence that pedestrian and bicycle crashes occurring on local roads form spatial patterns among certain sociodemographic groups. Road traffic crashes can be analyzed from different spatial contexts to establish spatial associations. The measurements of these spatial dependencies integrated with GIS can help analyze spatial patterns and clusters in crash data to improve traffic safety.

There is vast existing literature on traffic safety analysis. However, majority of traditional traffic safety studies evaluated the impacts of various attributes deterministically, either qualitatively or quantitatively, such as the influence of various geometric, demographic and environmental factors on crash occurrences [4]. The key assumption of these models is that factors leading to crash events are independent of each other. However, this assumption is often true for crashes occurring in a small area where geographical conditions can be assumed homogeneous and as such, this assumption could be violated for crash analysis based at County or State level. The presence of spatial dependencies where values at one location are influenced by presence of other values in its geographic proximity often violates the assumption of independence that is applied in many statistical analyses [5]. Therefore, failure to account for spatial dependencies may often result in erroneous predictions in statistical analysis for geographic data because features lying in space influenced by geographical factors are bound to display some sort of spatial dependencies [6]

GIS application has increased enormously because of its ability to measure spatial autocorrelation in data [3, 7, 8, 9]. For pedestrian crashes that are influenced by geographical factors, it is important to analyze the spatial dependences of crash data spread in space. Various

tools are available in GIS to analyze spatial association of data. Understanding when to use which tool, and why can be confusing [10]. Generally, identifying spatial patterns in GIS can be categorized into two groups depending on their output. The first category consists of global measures such as Ripley's K-function, Getis's G-statistic and Moran's I. These global measures simply test whether a given point distribution differs from a random distribution. They can examine if there exists a general tendency to clustering of crashes and they do not reveal the location of clusters within the distribution [11]. The second category consists of local measures such as kernel density and the local-autocorrelation methods, which identify exact position of a cluster within a section or within a network. The methods from second category are more efficient as they are concerned with spatial dependencies on a localized scale [11].

In addition to locating the exact location of clusters, kernel density also offers an advantage of determining the spread of risk of an accident by defining a search radius (also called bandwidth) around a defined cluster in which there is an increased likelihood for a crash to occur based on spatial dependency. Previous studies such as [7] identified critical areas with high child pedestrian crash risk using kernel density estimation. The study [3] created kernel density maps subsequently disaggregated it to create a basic spatial unit of an accident hotspot. Another study [12] developed a kernel estimation method to automatically identify road traffic accident hot spots in Christchurch in New Zealand.

However, kernel density has one major drawback, which is the failure to determine the statistical significance of resulting clusters [3]. Researchers however supplement kernel density with statistical techniques to identify significant clustering. For instance, [12] used kernel density in conjunction with Monte Carlo simulation techniques, to identify statistically significant clusters. The study [13] identified hot spots using kernel density and applied Poisson distribution to test the statistical significance of contributing factors.

2. METHODOLOGY

Shelby County is located farthest West of Tennessee state, covers a geographical area of 784 mi^2 and is the most populous County with a population of 927,644 people according to 2010 census. Yet, traffic safety facts indicate that Shelby ranks highest in pedestrian fatalities Statewide [14]. Crash records from Tennessee Department of Transportation (TDOT) indicate that the majority of pedestrian and bicycle crashes occurred in Shelby County (Figure 1) for the study period from 2008 to 2012.

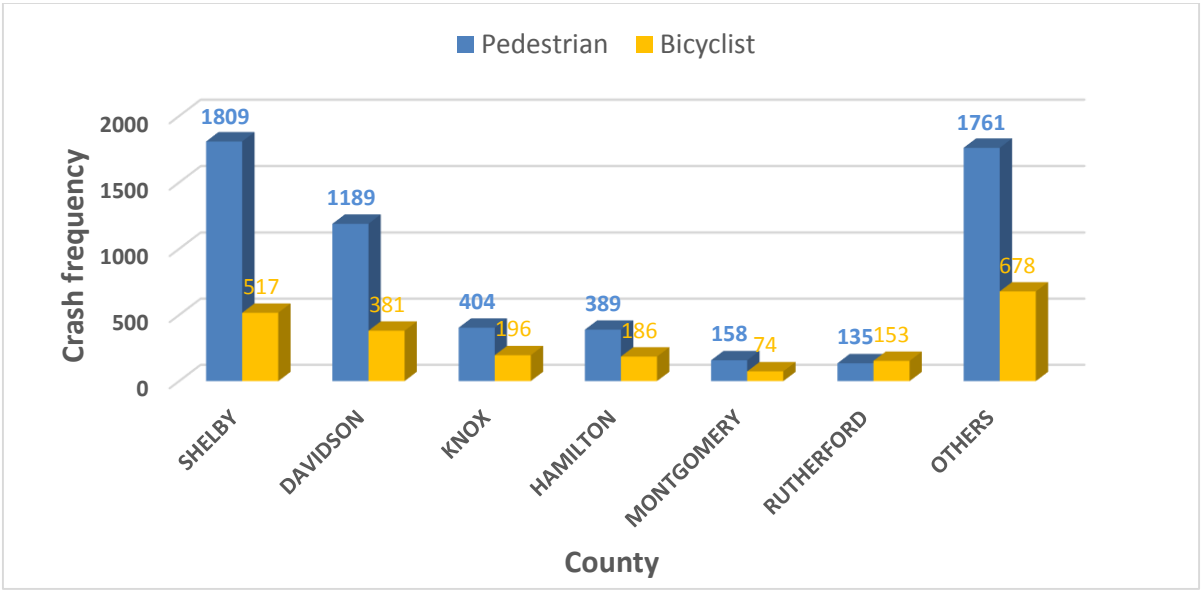

Figure 1: Total Pedestrian and bicycle crashes by County (2008-2012)

Meanwhile pedestrian crashes happen in Shelby because of high pedestrian exposures arising from high population density and different sociodemographic groups. Forexample the study [15] found pedestrian and bicycle crash clusters in major cities of Tennessee specifically Mephis in Shelby County and indicated that these crashes were correlated with percentage distribution of population by race, age groups, mean household income, percentage in the labor force, poverty level, and vehicle ownership.

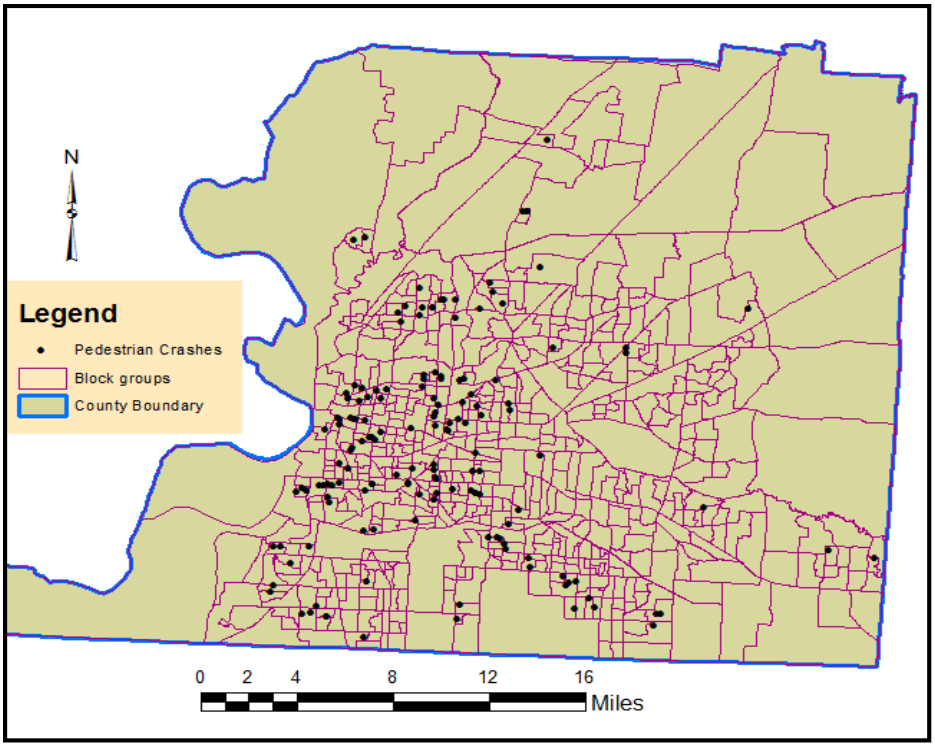

Figure 2: Map Shelby County showing distribution of local road pedestrian crashes

Hauer suggested that these traffic accidents could be reduced by applying appropriate remedial measures [16]. However, the success of these remedial measures depends on in-depth analysis of traffic accident records. Therefore, it is important to collect accurate, precise and reliable data with the traffic accident reports [13]. In this study pedestrian and bicycle, crash data were obtained from Tennessee Road Information Management System, a database managed by Tennessee Department of Transportation. Because of small number of bicycle crashes, this study focused only on pedestrian crashes for data analysis. Out of the initial 4,816 pedestrian crash records collected from 2008 to 2012, 492 crashes occurred on local roads in Tennessee. Sociodemographic data was obtained from US census bureau and was collected at Census block group level. The 2006-2010 American Community Survey 5-year estimates were used and contain 4,125 block groups with information such as population counts, household income, poverty status, population age, mode of transport to work and household vehicle ownership.

Basing on these data, a GIS kernel density methodology was developed to identify crash clusters by analyzing spatial patterns of crashes. Crash clusters define locations with unusual high concentration of crash occurrence. McCullough, suggests that many of the cluster analysis techniques use the base idea of computing the number of cases within a determined area and then testing the count that results for statistical significance using a Poisson or Bernoulli type statistical test [17]. This study based on two different methods to determine clustering of pedestrian crashes. First, we computed threshold value using Poisson distribution for Census block group crash counts. Poisson distribution was used to estimate the number of crashes that would be expected to occur in a block group in a given period. If Y is a random variable that describes accidents occurrence over time and y represents the observed number of accidents over a given time-period, then the mean of Y is Λ (also the random variable), and where $\Lambda = \mu$, Y is Poisson distributed with parameter μ.

The number of pedestrian crashes that occurred in a five-year period was calculated in each Census block group by conducting spatial analysis in Arc GIS and consequently we created a frequency table for observed crashes in each Census block group for the entire study period. The table lists number of block groups containing no accident, one accident, two accidents and so on. Then frequency table for the expected distribution was calculated based on Poisson distribution. To do this, the probability that a given number of accidents will occur in a Census block group is calculated first by finding the average number of accidents per block group, referenced as μ.

$$\mu = \frac{n}{N} \tag{1}$$

where, μ is average number of accidents per block group, n is the number of pedestrian crashes and N is the number of Census block groups. μ is used to determine the probability of particular of pedestrian crashes occurring in a given Census block group. The Poisson probability function can be written as follows,

$$p(x) = \frac{e^{-\mu}\,\mu^{x}}{x!} \tag{2}$$

where, *p(x)* is the probability of *x* number of pedestrian crashes per Census block group, *e* is Euler's constant. This equation is used to calculate the probability of crash occurrence, and these probabilities are listed in the frequency Table 1. The probability for each category of crashes (i.e. 0,1…4 or more) is multiplied by the total number of Census block groups to get the number of block groups expected to experience certain number of crashes. The expected and observed frequency distributions are compared to determine whether the accidents create a pattern. If the table for the observed distribution has more block groups containing many accidents than those of the table for the random distribution, then the crashes create a clustered pattern. However, Chi square test does not account for the location of each Census block group, but it is an appropriate test to make an initial comparison between both distributions. From Figure 3, it can be seen that there is about 88.8% chance that a local road pedestrian crash will not occur in a given Census block group.

Table 1: Observed and expected frequency distributions

Number of crashes (x)	Observed frequency (f_o)	Total number of crashes	Probability P(x)	Expected frequency (f_e)
0	3724	0	0.888	3661.21
1	328	328	0.106	436.68
2	58	116	0.006	26.04
3	12	36	0.000	1.04
4 or more	3	12	0.000	0.03
	4125	492	1.000	4125.00

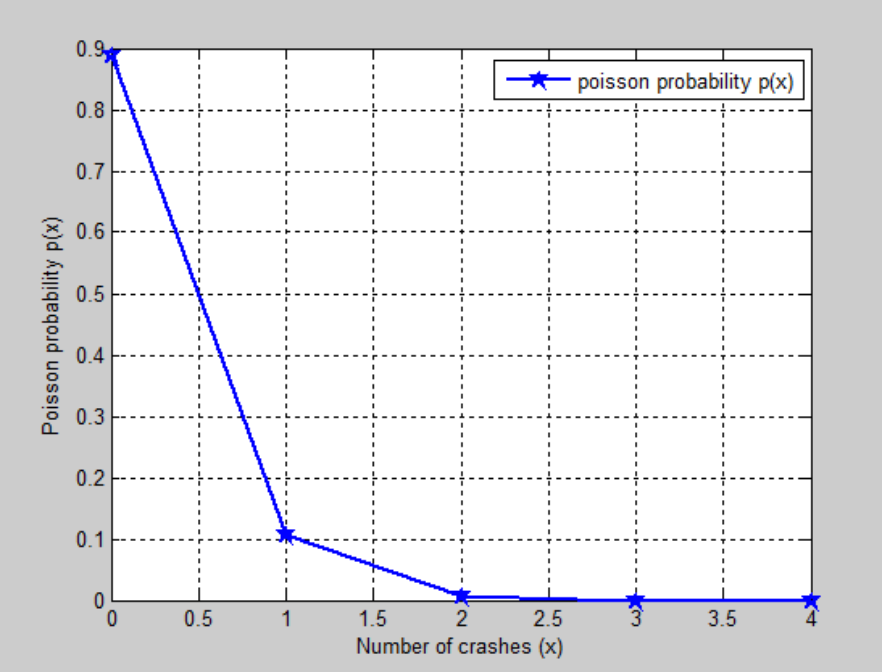

Figure 3: Poisson probability of crash occurrence

Chi-Square Analysis

The study tested goodness of fit of data to a Poisson distribution by comparing observed (actual) and expected frequencies. Chi-square test was used to determine whether two frequency distributions are significantly different. For each category of crashes, the expected number of Census block groups (f_e) is subtracted from the observed number of Census block groups (f_o). The difference is squared and divided by expected number and the results are then summed for all frequency levels. The Chi-square statistic (χ^2) is given by the following equation;

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(f_{o_i} - f_{e_i}\right)^2}{f_{e_i}} \quad (3)$$

f_{o_i} = observed frequency

f_{e_i} = expected frequency

k = number of frequency classes

The purpose of Chi Square test is to provide for one of two decisions: First, it is not very likely that the observed distribution is in fact identical with the expected distribution, and second, the observed distribution could be identical with the expected distribution. Therefore, it can be seen that either decision can be erroneously made. The first decision can be wrong if in fact the expected distribution is the observed distribution. On the other hand, second decision can be wrong if the observed distribution is in fact different from the expected distribution. Statistical tests of significance allow for specifying the probability of making either of these types of error. Common significance levels are (0.01, 0.05 and 0.10). For instance, when a test is made at the 5% confidence level, the engineer takes the chance that 5%, of the rejected expected distributions are in fact identical with the corresponding observed distributions.

Table 2: Chi Square distribution

x	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
0	3724	3661.21	62.791	3942.77	1.08
1	328	436.68	-108.682	11811.85	27.05
2	58	26.04	31.958	1021.30	39.22
3	12	1.04	10.965	120.22	116.12
4+	3	0.03	2.969	8.82	285.55
					469.01

To determine whether the number of crashes follows a Poisson distribution, the null and alternative hypotheses are:

Null hypothesis, H_0: Observed and expected distributions are identical.

Alternative hypothesis, H_1: Observed and expected distributions are significantly different.

The Poisson distribution has one parameter, its mean μ, whose value is required in the null and alternative hypotheses. The mean can be either a value based on past knowledge, or a

value estimated from sample data. In this study, the mean of crashes was estimated from the data as 0.12. Because we estimated the mean of the Poisson distribution from the data, the number of degrees of freedom are = 5 − 1 − 1 = 3. Using the threshold value of 0.05 level of significance, the critical value of χ^2 with 3 degrees of freedom is 7.815 (from Chi Square distribution table). The decision rule is; Reject H_0 if $\chi 2$ > threshold value; otherwise do not reject H_0. From Table 2, since $\chi 2 = 469.01 > 7.815$, the decision is to reject H_0. Therefore, there is sufficient evidence to conclude that the data fits a Poisson distribution and therefore we were 95% confident that there is significant difference between the observed and the expected data. The distribution curve for observed and expected frequencies can be seen in Figure 4.

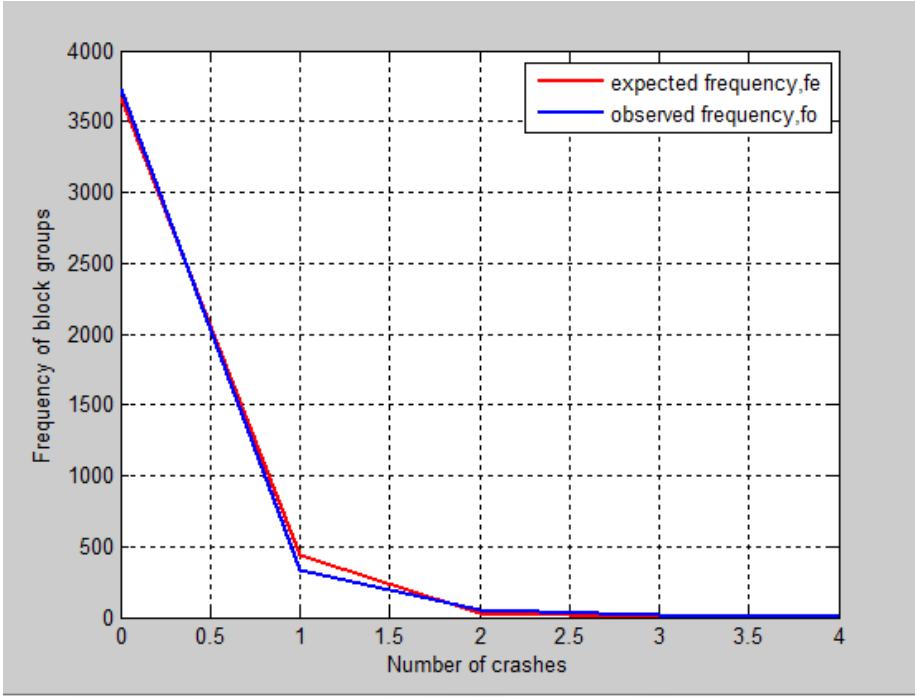

Figure 4: Observed and expected distribution of crashes

After establishing that crashes create a clustered pattern in block groups, kernel density method in GIS was used to identify the location of crash clusters in a search radius of 0.5-mile area. Kernel Density calculates a magnitude per unit area from point or line features using a kernel function to fit a smoothly tapered surface to each point or line. Kernel density estimation involves placing a symmetrical surface over each point and then evaluating the distance from the point to a reference location based on a mathematical function and then summing the value for all the surfaces for that reference location as shown in Figure 5. This procedure is repeated for successive points and allows placing a kernel over each crash observation. By summing these individual kernels to give the density estimate for the distribution of crash points [3]. The kernel function can be expressed as;

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1} K\left(\frac{d_i}{h}\right) \qquad (4)$$

Where $f(x, y)$ is the density estimate at the location (x, y), n the number of observations, h is the search radius or bandwidth, K is kernel function, and d_i is the distance between the location (x, y)

and the location of the i^{th} observation of features that fall within the search area. The search radius affects the resulting density map in a way that if the radius is increased there is a possibility that the circular neighborhood would include more feature points which results in a smoother density surface [9, 18]

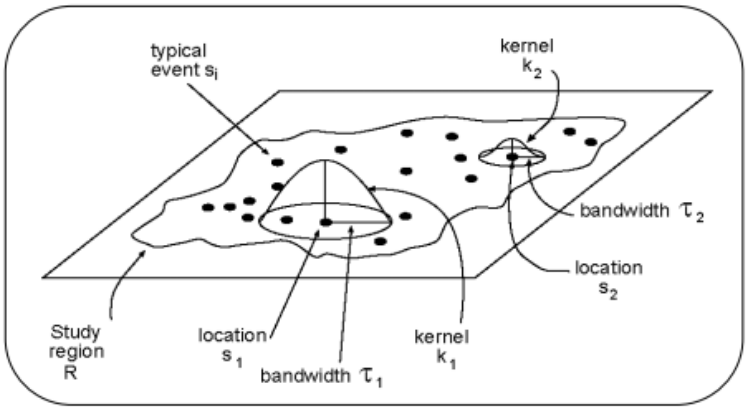

Figure 5: Principle of kernel density (Source: Erdogan, et al. 2008)

3. RESULTS AND DISCUSSION

Kernel density tool was used to create crash density maps using a search radius of 0.5 miles. The resulting density map is presented in Figure 6. After determining high crash zones (clusters), we investigated the accident causal factors at these locations. Sociodemographic data was overlaid on the cluster map to identify neighborhood characteristics of each pedestrian-vehicle crash hot spot. We determined association of crash clusters with respect to average household income, poverty status, vehicle ownership, education attainment and mode of transport to work. Results in Figure 8 indicate that clustering is evident among households below poverty level. Poverty threshold is defined by US census bureau basing on the total household income and the number of family members [19]. Similarly, crash clusters were observed among neighborhoods of low average household income (Figure 7). Low-income populations travel less frequently, have the lowest income groups and are much less likely to own an automobile [20]. Lower mobility of low-income households may reflect their higher rates of unemployment, lower education attainment, and low vehicle ownership. Moreover, results from this study indicate clustering among neighborhoods of households without vehicles (Figure 9), implying more people make their trips by foot compared to households that own 2 or more cars, which expose them to crash risks. Additionally, car ownership is highly associated with travel mode choice, for instance clusters were observed among block groups with large number of people who commute to work by walking whereas clusters were not evident among neighborhoods with population that commute to work by private car. This study also identified a correlation between education and pedestrian crashes. For example, crash clusters were observed in areas with high number of people with education less than high school.

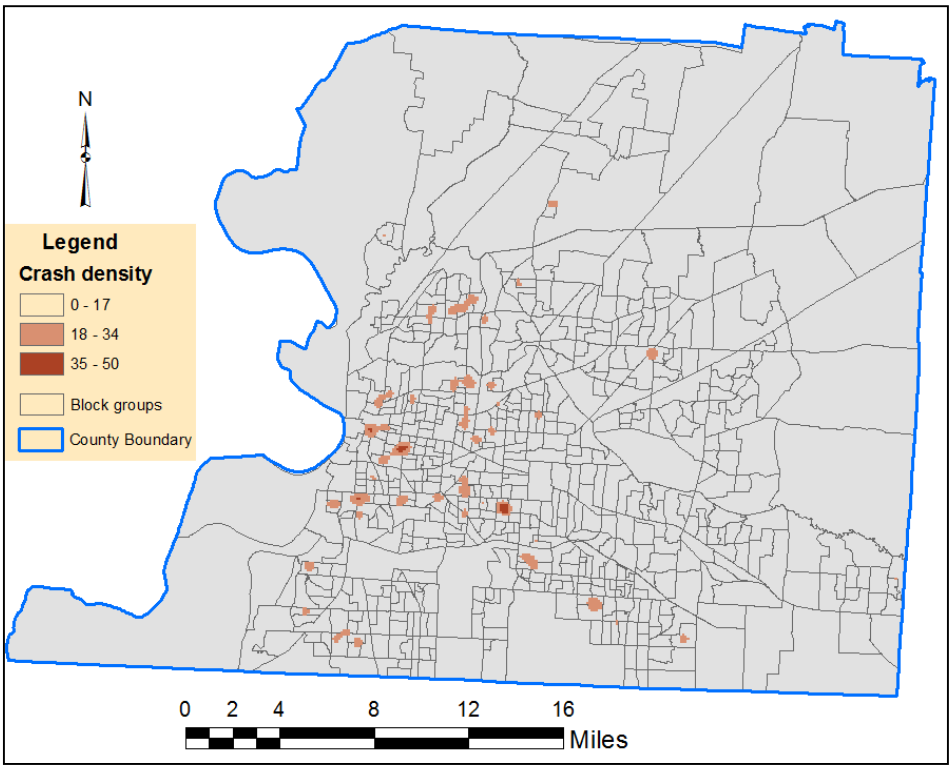

Figure 6: Map showing location of clusters

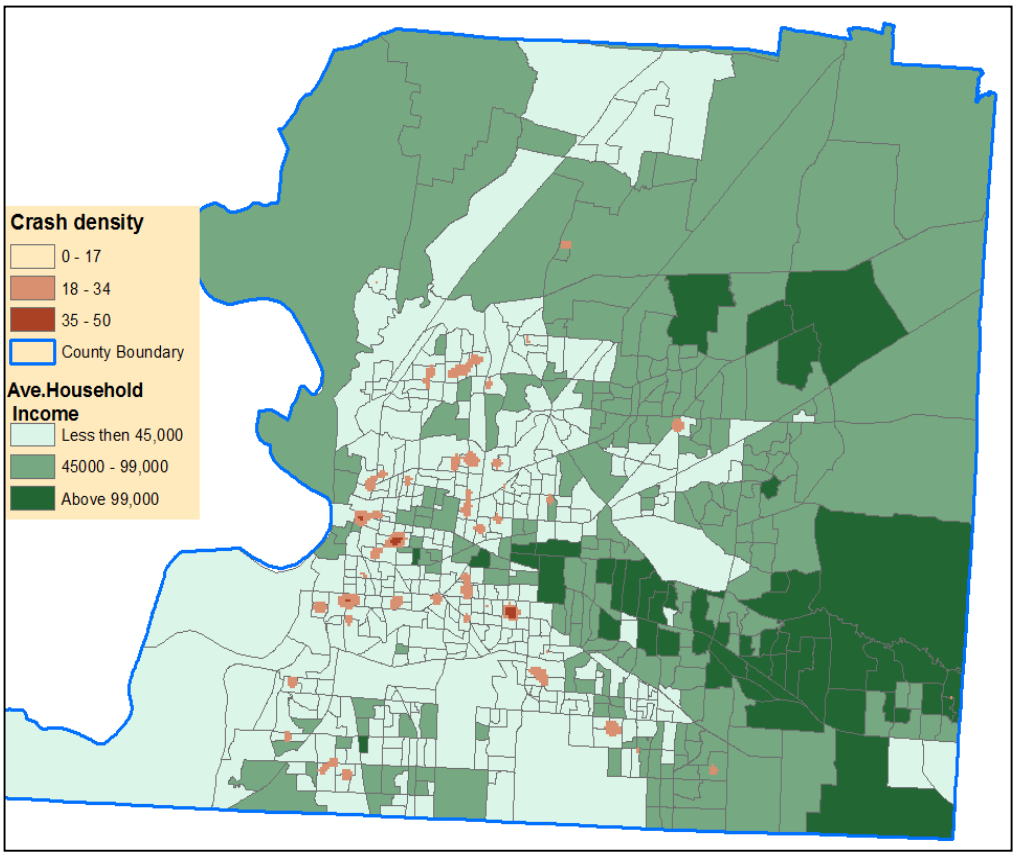

Figure 7: Pedestrian crash clusters with average household income

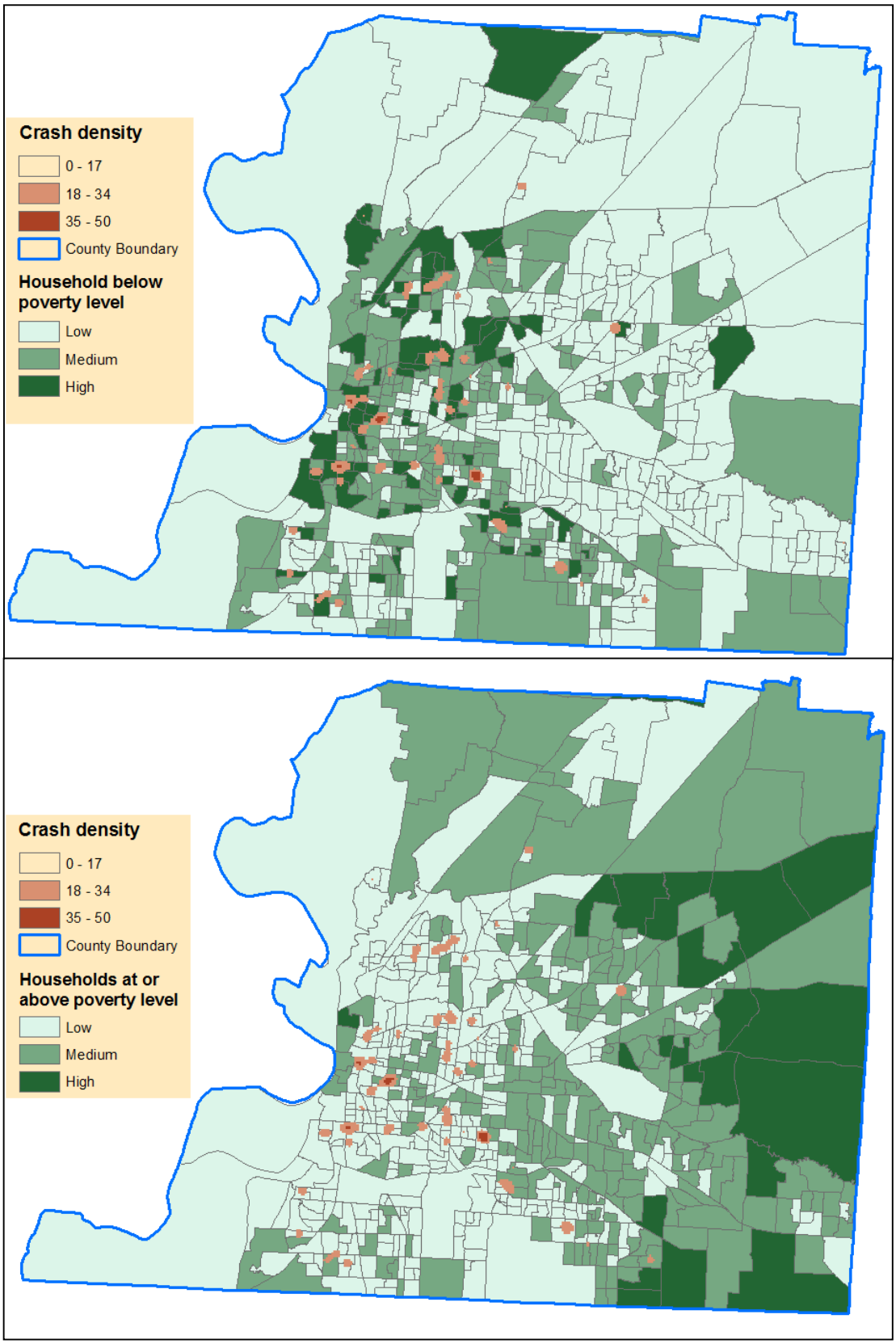

Figure 8: Pedestrian crash clusters with poverty status

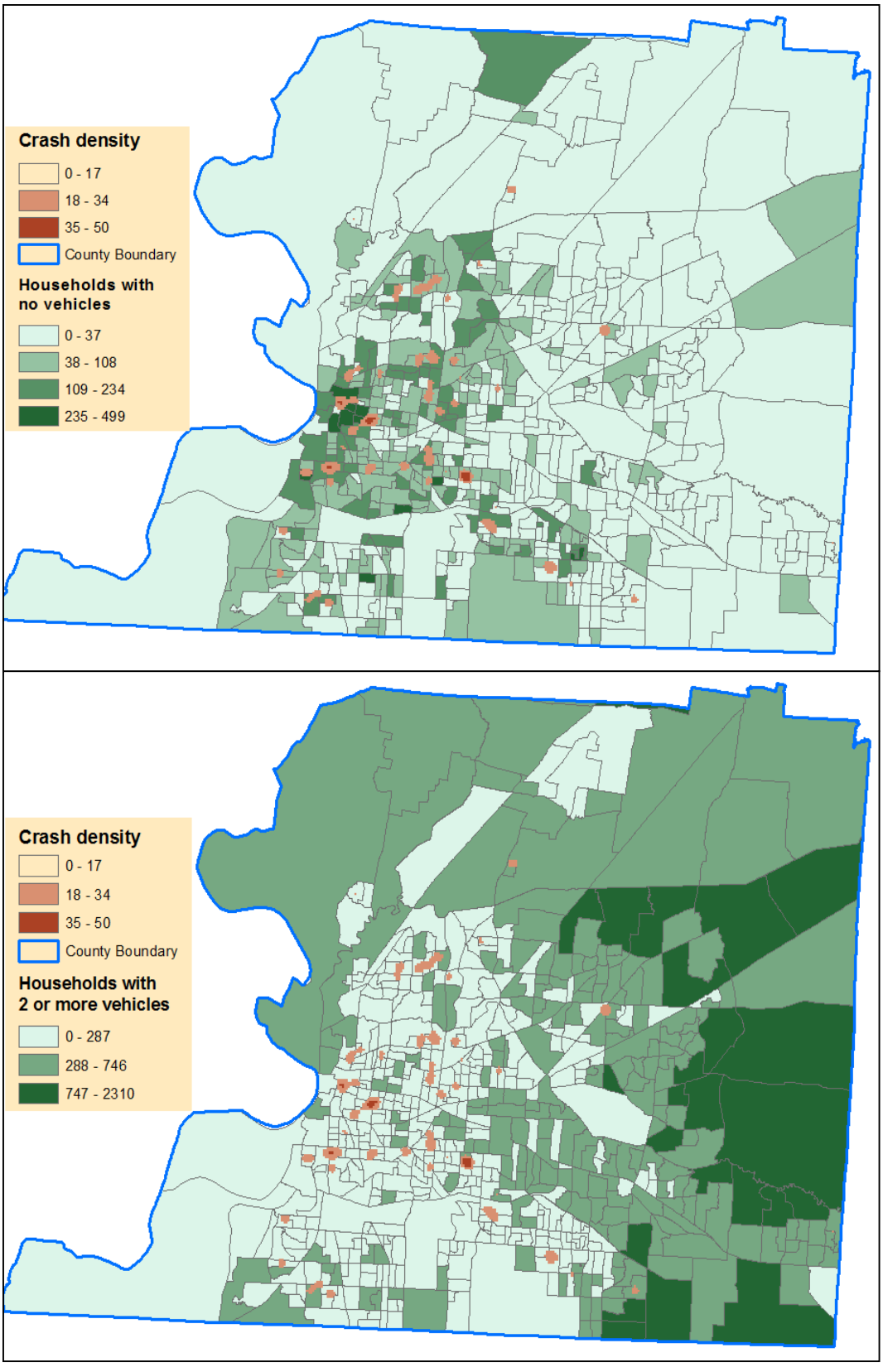

Figure 9: Pedestrian crash clusters with household vehicle ownership

4. CONCLUSIONS

This paper presents a GIS methodology to identify high-density local road pedestrian crash zones and determines the causal factors. The use of this methodology is based on statewide crash data collected in Tennessee from 2008 to 2012. At first Poisson distribution was developed to ascertain whether crashes create clustered pattern by comparing the frequency distribution of observed Census block groups containing many rashes and those of expected (random) distribution. Chi square analysis was performed to find out whether two distributions are significantly different. The GIS kernel density estimation tool was then used to create crash density map and locate crash clusters. By overlaying socioeconomic data on density map, crash clusters were evidently associated with low-income households, households below poverty level, population with education less than high school and households without vehicles. Our study helps to explain why pedestrian crashes are more frequent with certain sociodemographic groups than with others. These results are useful to guide traffic planning process and can assist local decision-makers to develop effective countermeasures to reduce pedestrian crashes.

References

[1] W. W. Hunter, W. E. Pein and J. C. Stutts, "Pedestrian and Bicycle Crash Types of The Early 1900's," Federal Highway Administration,Publication No. FHWA-RD-95-163, McLean,VG, 1996.

[2] FHWA, "GIS Tools for Improving Pedestrian & Bicycle Safety TechBrief, FHWA RD-00-153, Federal Highway Administration (FHWA)," U.S. Department of Transportation (U.S. DOT), 2000.

[3] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis and Prevention ,* vol. 41, p. 359–364, 2009.

[4] V. Shankar, F. Mannering and W. Barfield, "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies," *Accid. Anal. and Prev,* vol. 27, no. 3, p. 371–389, 1995.

[5] G. Khan, X. Qin and D. A. Noyce, "Spatial Analysis of Weather Crash Patterns," *Journal of Transportation Engineering,* vol. 134, no. 5, pp. 191-202, 2008.

[6] A. Getis and J. K. Ord, "The Analysis of Spatial Association by Use of Distance Statistics," *Geographical Analysis,* vol. 24, no. 3, p. , 1992.

[7] C. A. Blazquez and M. S. Celis, "A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile," *Accident Analysis and Prevention,* p. 304–311, 2013.

[8] A. Bulajića, D. Jovanovićb, B. Matovićc and S. Bačkalićd, "Identification of High-Density Locations with Homogeneous attributes of Pedestrian crashes in the Urban area of Novi Sad," Borsko Jezero, 2014.

[9] B. P. Loo, S. Yao and J. Wu, "Spatial Point Analysis of Road Crashes in Shanghai A GIS-based Network Kernel Density Method," in *Geoinformatics, 2011 19th International Conference* , Shanghai, 2011.

[10] D. W. Allen, GIS Tutorial 2:Spatial Analysis Workbook, ESRI Press, 2001.

[11] M. Bíl, R. Andrásik and Z. Janoska, "Identification ofhazardous road locations oftraffic accidents by means ofkernel density estimation and cluster significance evaluation," *Accident Analysis and Prevention ,* vol. 55, p. 265–273, 2013.

[12] C. E. Sabel, S. Kingham, A. Nicholson and P. Bartie, "Road traffic accident simulation modeling-a Kernel estimation approach.Presented at SIRC The 17th Annual Colloquium of

the Spatial Information Research Centre University of Otago," Dunedin, New Zealand, 2005.

[13] S. Erdogan, I. Yilmaz, T. Baybura and M. Gullu, "Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar," *Accident Analysis and Prevention* , vol. 40, p. 174–181, 2008.

[14] National Highway Traffic Safety Adminitration, "Traffic Safety Facts for Tennessee : 2008-2012," US Department of Transportation, Washington,DC, 2012.

[15] D. Chimba, D. Emaasit, C. R. Cherry and Z. Pannell, "Patterning Demographic and Socioeconomic Characteristics Affecting Pedestrian and Bicycle Crash Frequency," in *Transportation Research Board*, Washington, 2014.

[16] E. Hauer, "Identification of Sites with Promise," in *Transportation Research,74th Annual Meeting*, Washington, Dc, 1996.

[17] M. J. McCullagh, "Detecting hotspots in time and space. International Symposium & Exhibition on Geoinformation 2006," SubangJaya, Selangor, Malaysia, 2006.

[18] S. S. Pulugurtha, V. K. Krishnakumar and S. S. Nambisan, "New methods to identify and rank high pedestriancrash zones: An illustration," *Accident Analysis and Prevention*, vol. 39, p. 800–811, 2007.

[19] S. Srinivasan and K. Jermprapai, "A Planning-Level Model for Assessing Pedestrian Safety," in *Transportation Research Record*, Washington,DC, 2014.

[20] J. Pucher and J. L. Renne, "Socioeconomics of Urban Travel: Evidence from the 2001 NHTS," Eno Transportation Foundation, Inc, Washington, DC, 2006.