EVALUATION OF HOTSPOTS IDENTIFICATION USING KERNEL DENSITY ESTIMATION (K) AND GETIS-ORD (G_i*) ON I-630

Uday R. R. Manepalli
Graduate Student, Civil, Architectural and Environmental Engineering,
Missouri University of Science & Technology
Rolla, MO, USA, E-mail: umr75@mail.mst.edu

Ghulam H. Bham*, Ph.D
(Corresponding author)
Assistant Professor, Civil, Architectural and Environmental Engineering,
Missouri University of Science & Technology
Rolla, MO, USA, E-mail: ghbham@mst.edu

Srinadh Kandada
Graduate Student, Civil, Architectural and Environmental Engineering,
Missouri University of Science & Technology
Rolla, MO, USA, E-mail: skp6d@mail.mst.edu

Submitted to the 3rd International Conference on Road Safety and Simulation,

September 14-16, 2011, Indianapolis, USA

No. of words in abstract =248

No. of words in text = 4245 + 1750 (1 table and 6 Figures) = 5995

ABSTRACT

The main objective of this paper is to compare two statistical techniques kernel density estimation (K) and Getis-Ord G_i* statistic using a Geographic Information System (GIS) for hotspot identification. The standardized G_i* is essentially a Z-value associated with statistical significance. The two statistical techniques were compared using seven years of crash data (2000-2006) on I-630 (7.4 miles) in Arkansas. The highway had very high rate of crashes; 457 crashes per mile were observed during the analysis period. I-630 is located in only one county and, therefore, assumed that the demographic effect will be minimal and mainly local spatial autocorrelation will be observed. Results indicated that the estimation by K and G_i* including three conceptualization of spatial relationships (CSR) for hotspots (high; categorized as high and low) were almost similar. Additionally, the three CSR methods (fixed distance, inverse distance, and inverse square distance) identified the same hotspots (high). Also, the range of Z values for G_i* for the hotspots (high) were similar for the three CSR's for years 2000 and 2001. For 2002 and 2003, range of Z values for G_i* for the hotspots (high) were similar for inverse distance and inverse square distance CSR. Another data set, aggregated from 2004 to 2006 was used to identify hotspots by K and G_i* (inverse square distance, CSR). Results indicated similar hotspots identified by both methods. The reason may be due to the mathematical functions of K and G^*_i. Some of the key contributing factors in this paper are also discussed.

INTRODUCTION

"Hotspots," "black spots," or high crash locations are sites on a section of a highway that have an accident frequency significantly higher than expected at some threshold level of significance (Hakkert and Mahalel, 1978). The estimated highway crash cost to society in 2000 was a staggering $230 billion a year according to NHTSA, 2006. To reduce the highway crash cost, a plausible solution is the accurate identification of hotspots. Some of the commonly used hotspot identification methods are empirical Bayes (EB) (Hauer et. al., 2002), crash rate (CR) (Powers and Carson, 2004), and equivalent property damage only (EPDO) (Campbell and Knapp, 2005). Among the existing methods, EB method is considered to the best hotspot identification method. However, many cities and departments of transportation (DOTs) still use crash counts and rates to identify and rank high crash segments (Mitra, 2009). The crash rate and counts are simpler and straightforward to use, but not superior than the EB method. For methods like the EB method, special training and skills in statistical analysis are required; one of the reasons why DOTs still rely on simple methods (Mitra, 2009).

Past studies have used crash data to identify high spatial concentrations of crashes (Norden et. al., 1956; Hakkert and Mahalel, 1978; McGuigan, 1991; Depue, 2003; Songchitruksa and Zeng, 2010) using GIS. Though GIS-based methods may or may not be as superior as the EB method, they are at least better than methods that use crash counts or crash rates. Moreover, when overlaid with other layers, GIS-based mapping could help to associate high-crash locations with spatial factors. GIS have been widely used to geocode accident locations and develop maps of crashes using database queries (Levine et. al., 1995a; Levine et. al., 1995b; Affum and Taylor, 1995; Kin and Levine, 1996; Austin et. al., 1997; Miller, 1999). McMohan (McMohan, 1999) used GIS to analyze pedestrian crash risk using buffering, cluster analysis, and spatial queries. Peled et al. (1996) used a GIS to generate maps of the distribution of crash concentrations. Some of the other methods used are kernel density estimation (Flahaut et. al., 2003; Pulugurtha et. al., 2007) and local spatial autocorrelation (Flahaut et. al., 2003) to identify hotspots.

Point pattern analyses have been widely examined by scientists and a variety of methods were developed for detecting hotspots. The point pattern methods can be classified broadly into two categories (O'Sullivan and Unwin, 2002, i) methods which examine first-order effects, which measure the variation in the mean value of the process like kernel density estimation, quadrant count analysis etc, and ii) methods which examine second-order effects which measure the spatial dependency of points for spatial patterns like Moran's I, Getis-Ord G statistic (Xie and Yan, 2008).

The main objective of this paper is to compare two statistical techniques, kernel density estimation (K) and Getis-Ord G_i* (Getis and Ord, 1992; Ord and Getis, 1995) statistics using a GIS for hotspot identification. The choice of choosing K and G_i* statistic is discussed in the later sections of the paper. Additionally, three conceptualization of spatial relationships (CSR) were used to determine G_i* namely fixed distance band, inverse distance and inverse square distance and the best CSR for hotspot identification is recommended. Also, this paper identifies that though these two methods have different conceptualizations, both produce similar results under specific conditions of the selected parameters. Further investigations into these methods (K and G*i), when the mathematical forms are considered, shows a relationship that leads to similar identification of hotspots. Some of the contributing factors for the occurrence of crashes are considered and explained further in the results.

The data and the choice of the methods used are explained next. Followed by the spatial autocorrelation index Getis-Ord G_i* and kernel density estimation (K) and the methodology. Later the results are explained and the paper ends with conclusions and recommendations for future research.

DATA USED AND CHOICE OF METHODS

Crash data for seven years (2000-2006) on I-630 in Arkansas have been used in this study, provided by the Arkansas Highway and Transportation Department (AHTD). I-630 is 7.4 miles in length and is located in only one county. I-630 has an average of 457 crashes per mile during the analysis period, AADT with an average of 94,921 vehicles/year. I-630 is located in Pulaski County, where the state capital of Little Rock is located. Little Rock has a population of 183,133 and an area of 116.8 square miles (US Census, 2008).Other major highways like I-40, I-30, I-430, I-440 and I-530 are also located in Pulaski County. Figure 1, illustrates the location of I-630 and the several highways surrounding it. It is assumed that the demographic effect will be limited to local conditions and may not vary widely in space. Therefore, local spatial autocorrelation, Getis-Ord G^*_i, is preferred for hotspot identification. Further specific details of G_i* statistic are discussed in the next section of the paper. The purpose of K is to produce a smooth density surface of point events (crashes) over space by computing event intensity as density estimation (Xie and Yan, 2008). K is one of the most popular methods for identification of hotspots, as it is easy to understand and implement (Bailey and Gatrell, 1995; Silverman, 1986). Further, specific details of K are discussed in the next section of the paper.

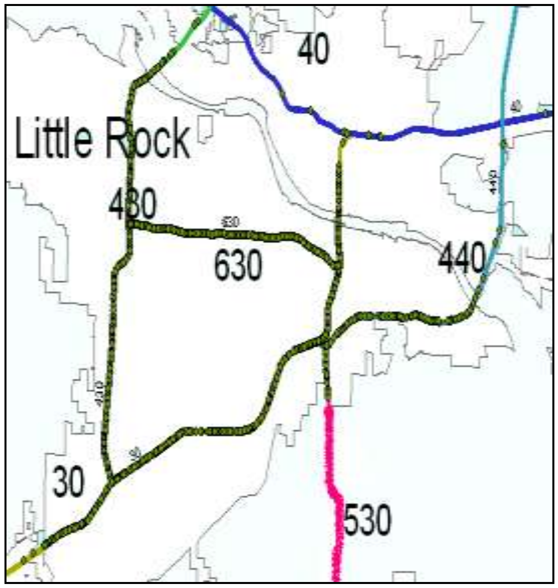


Figure 1. Location of I-630

Spatial autocorrelation

The first law of geography states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). The basic principle of spatial autocorrelation (SA) is similar and is defined as the correlation of a variable with itself through space. SA measures the strength of spatial autocorrelation and tests the assumption of independence (or randomness). If there are any systematic patterns in the spatial distribution of a variable, that variable is said to be spatially autocorrelated. If nearby areas are alike, the SA is positive. Negative autocorrelation applies to neighboring areas that are unlike, and random patterns exhibit no SA. SA also permits the examination of co-variations in properties observed in a two-dimensional geo-surface. SA indices do not explain why locations that have clusters of statistically significant crashes have higher incidence of crashes than other locations; therefore, SA methods cannot identify factors that cause crashes (Mitra, 2009).

Getis-Ord G_i* Statistic

G-statistics, developed by Getis and Ord, analyze evidence of spatial patterns (Getis and Ord, 1992; Ord and Getis, 1995). They represent a global SA index. The G_i* statistic, on the other hand, is a local SA index. It is more suitable for discerning cluster structures of high or low concentration. A simple form of the G_i* statistic is (Songchitruksa, 2010):

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} x_j} \qquad (1)$$

where: G_i* is the SA statistic of an event i over n events. The term, x_j, characterizes the magnitude of the variable x at events j over all n; the CSI at a particular location. The distribution of the G_i* statistic is normal when normality is observed in the underlying distribution of the variable *x*. The threshold distance (i.e., the proximity of one crash to another) in this study was set to zero to indicate that all features were considered neighbors of all other features. This threshold was applied over the entire region of the study. The standardized G_i* is essentially a Z-value and can be associated with statistical significance as:

$$Z(G_i^*) = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{x} \sum_{j=1}^{n} w_{ij}^2}{s \sqrt{\dfrac{n \sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}} \tag{2}$$

Positive and negative G_i* statistic with high absolute values correspond to clusters of crashes with high- and low-value events, respectively. A G_i* close to zero implies a random distribution of events.

Kernel Density Estimation (K)

The kernel density method is a non-parametric method that uses a density estimation technique. It enables the observer to evaluate the local probability accident occurrence and degree of danger of a zone. For a given set of observations from an unknown probability density function, the Kernel estimator can be defined as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3}$$

where: *h* is called the smoothing parameter or bandwidth, *K* is called the kernel and $\hat{f}$ is the estimator of the probability density function f. Thus, the kernel estimator depends on bandwidth (h) and kernel density (K). For a given kernel *K*, the kernel estimator critically depends on the choice of the smoothing parameter h. An appropriate choice of the smoothing parameter should be determined by the purpose of the estimate.

Categorization of the Hotspots

The hotspots were categorized based on the Z-value of the G_i* statistic. Categorization can based be on six classification schemes: equal interval, defined interval, quartile, natural breaks, geometric interval, and standard deviation. In the natural breaks scheme, the classes are based on natural categorizing inherent in the data. The break points are identified by the class breaks that best group similar values and maximize the differences between the classes. The features are divided into classes of the boundaries which correspond to relatively big jumps in the data values. This classification scheme was best suited to the present study (Krygier, 2010; ArcGIS).

In this paper, hotspots were categorized using natural breaks and based on the Jenks' algorithm. This algorithm is a common method of classifying data in a choropleth map, a type of thematic map that uses shading to represent classes of a feature associated with specific areas (e.g., a population density map). This algorithm generates a series of values that best represent the actual breaks observed in the data, as opposed to some arbitrary classificatory scheme; thus, it preserves true clustering of data values. Further, the algorithm creates *k* classes so that the variance within categories is minimized (Lewis, 1996). In this study, the categorization was carried out in two categories i.e., high and low.

METHODOLOGY

Initially, four years of crash data (2000-2003) was used to identify the hotspots using G_i* and K density estimation. Three conceptualization of spatial relationships (CSR) of the G_i* statistic namely; fixed distance, inverse distance and inverse square distance were used. The results were obtained for each year and matched with the hotspots identified by K. The differences in the identification of hotspots based on these three CSR were studied. Later, crash data aggregated from 2004 to 2006 were used to determine the hotspots using inverse square distance for the G_i* statistic and K. Results were compared in both cases. To understand the similar identification of hotspots using these methods, some of the crash contributing factors were considered. Further, the mathematical forms of the two methods (K and G^*_i), were considered to identify if a relationship exists between them.

RESULTS

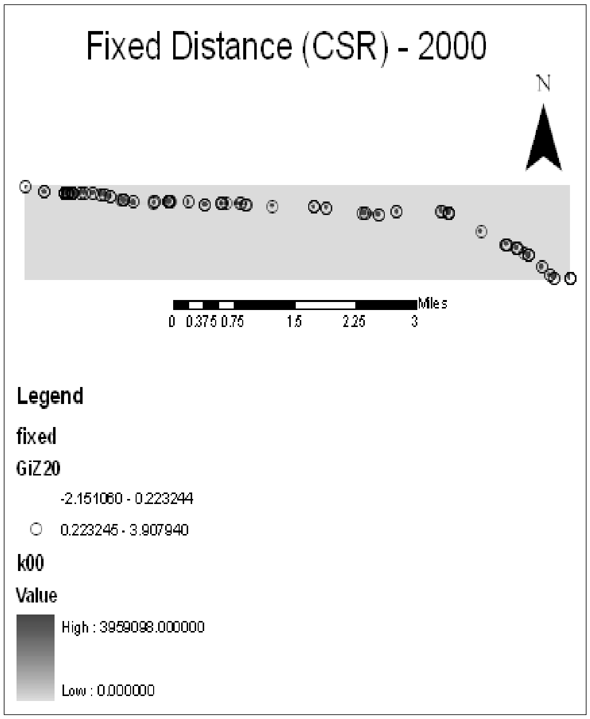
This section presents the description of crashes which occurred during the analysis period. Next, the results of the identification of hotspots by the three CSR for G_i* statistic and K are presented and discussed. This is followed by examining the mathematical forms of the two methods (K and G^*_i) to understand the reason for identification of similar hotspots. Finally, this section ends with presentation of some of the key reasons for crash contributing factors.

Table 1 presents the summary statistics of the crash frequency by year for I-630 from 2000 to 2006. Some of the major contributing factors are presented in terms of the percentage of crashes. From Table 1, for column 4, the remaining percentage of crashes represents straight roadway alignment i.e., for year 2000 among 500 crashes, 15% occurred on curve and 85% occurred on straight roadway profile. Similarly for column 5, the other factors include level roadway profile and unknown.

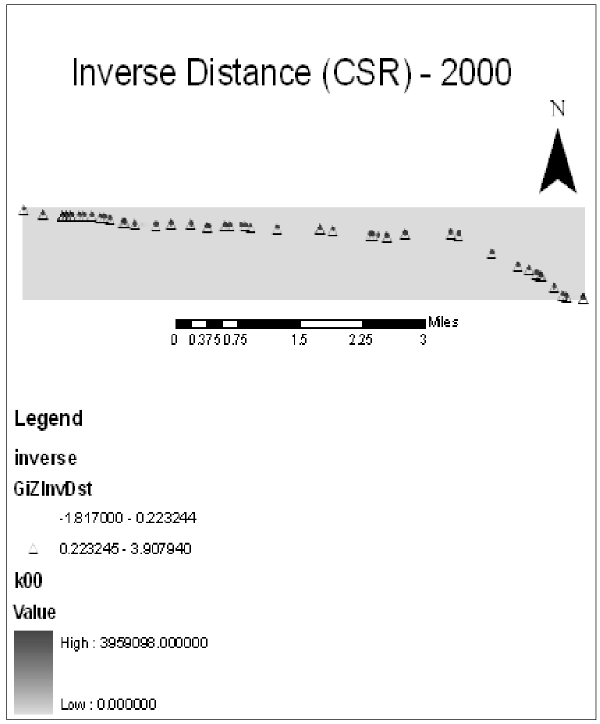
Table 1. Summary of crash statistics

Year (1)	CF (2)	AADT (vehs/yr) (3)	Curve* (4)	Grade* (5)	Types of Collisions* (6)			Weekends* (7)	DUI-Yes* (8)
					Rear-end (a)	Sideswipe Same Direction(b)	Single Vehicle Crashes (c)		
2000	500	90563	15	19	50	10	23	19	4
2001	523	93073	13	31	55	14	18	15	9
2002	443	93528	12	27	49	16	23	12	5
2003	537	97250	10	20	60	14	19	15	5
2004	423	98667	8	22	55	14	26	17	5
2005	446	90833	12	22	56	17	17	11	5
2006	511	100533	12	19	53	13	23	18	4

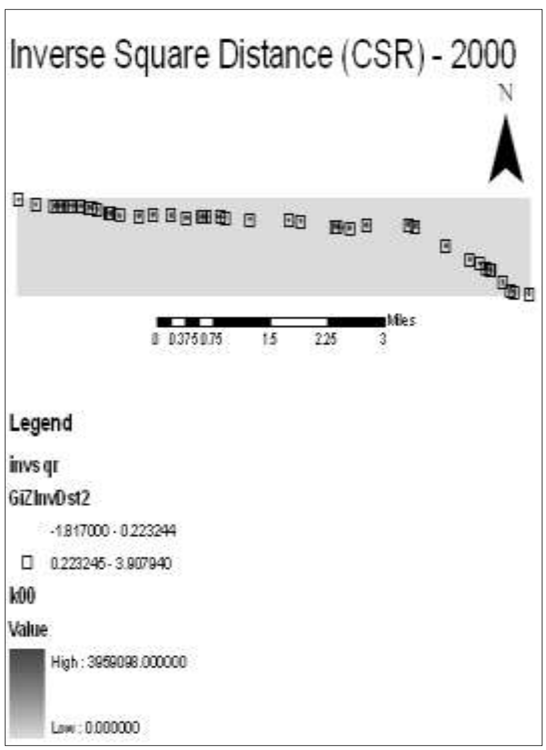
* indicate the percentage of crashes



(a)

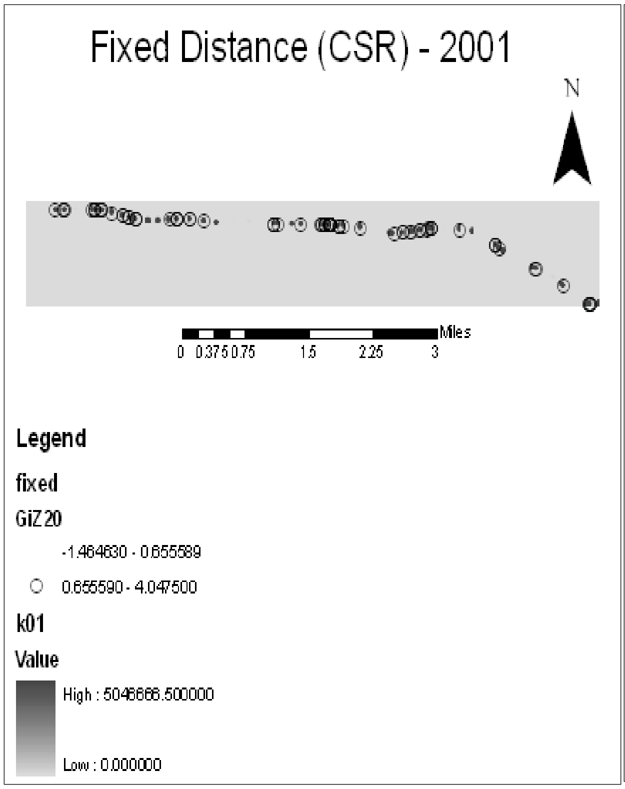


(b)

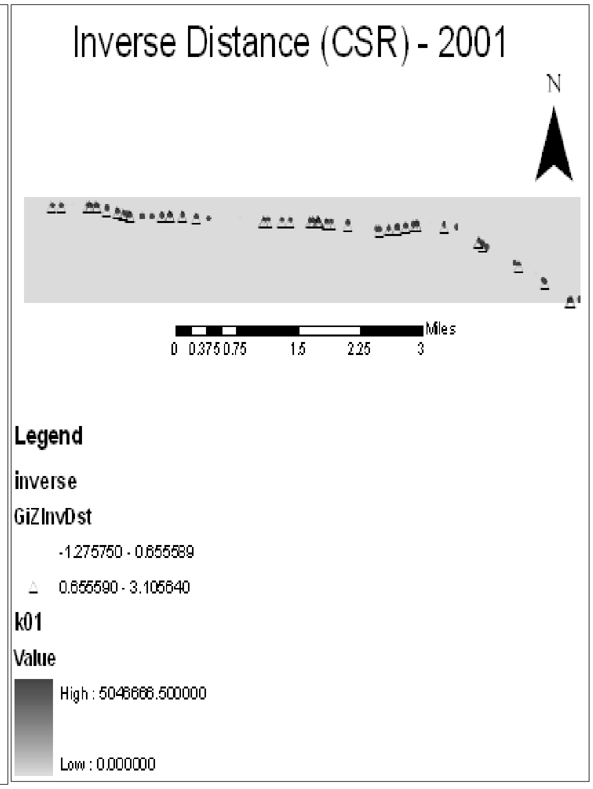


(c)

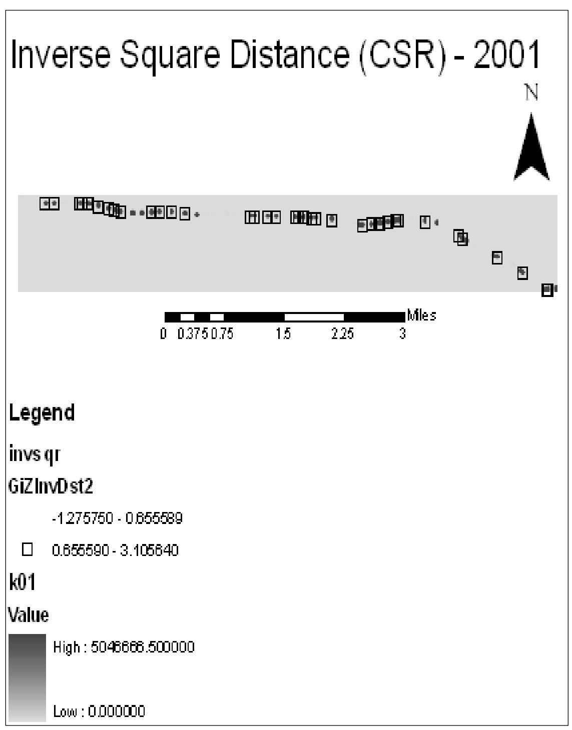
Figure 2. Hotspot identification by three CSR for G_i* statistic and K for year 2000



(a)

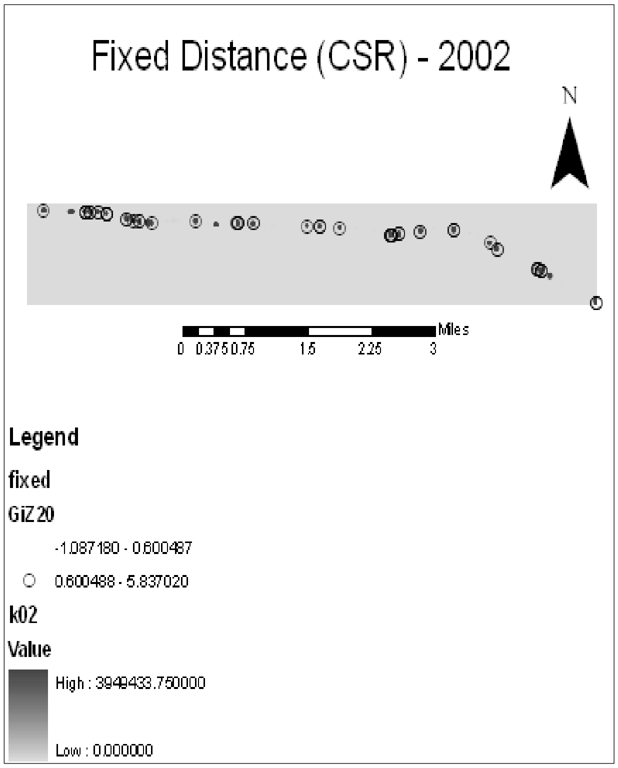


(b)

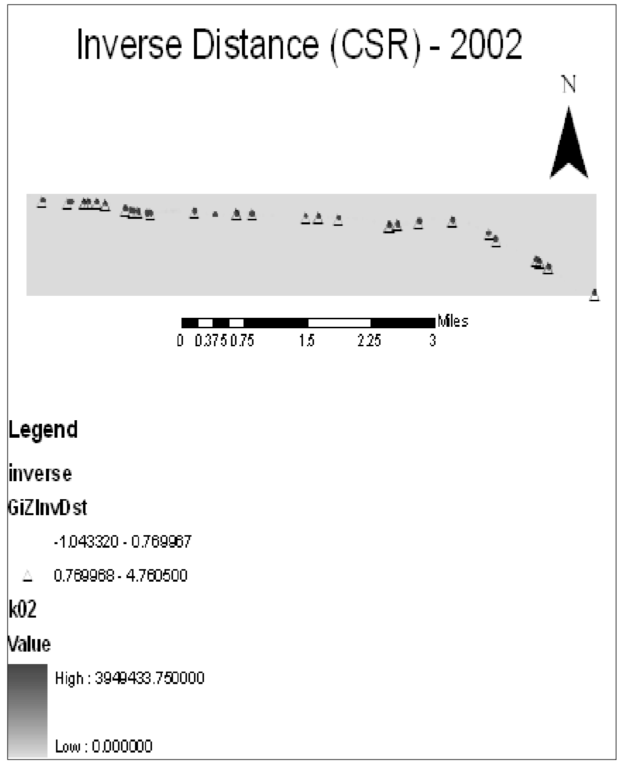


(c)

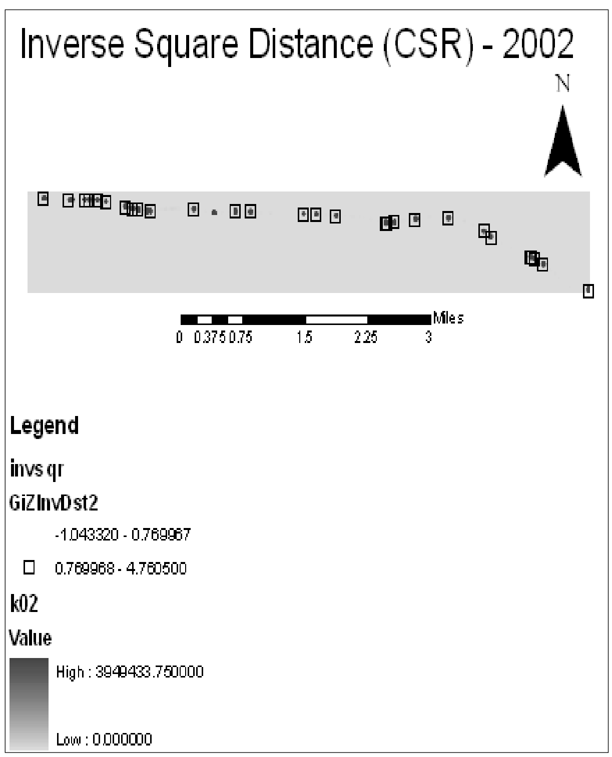
Figure 3. Hotspot identification by three CSR for G_i* statistic and K for year 2001



(a)

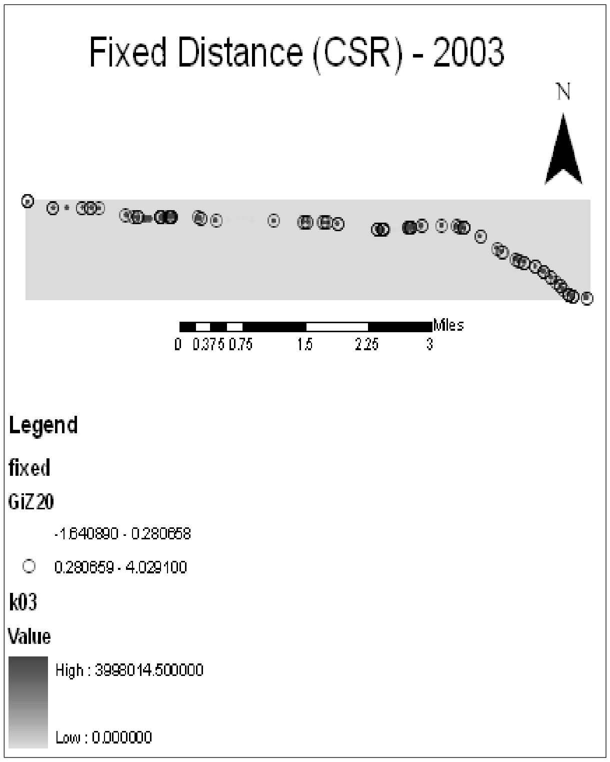


(b)

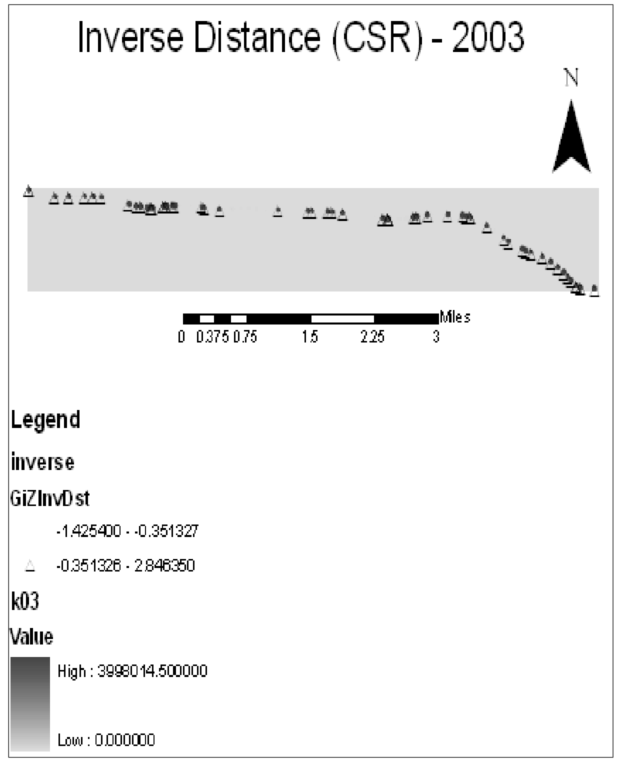


(c)

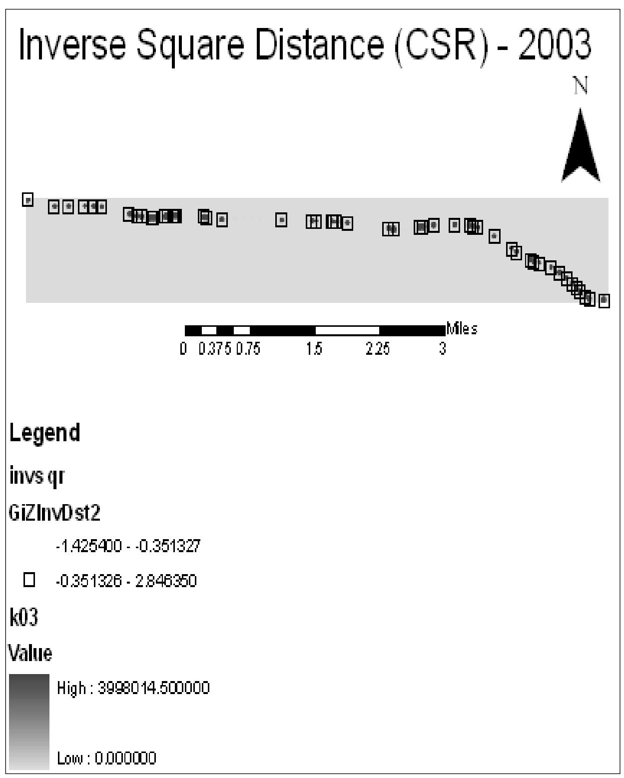
Figure 4. Hotspot identification by three CSR for G_i* statistic and K for year 2002



(a)



(b)



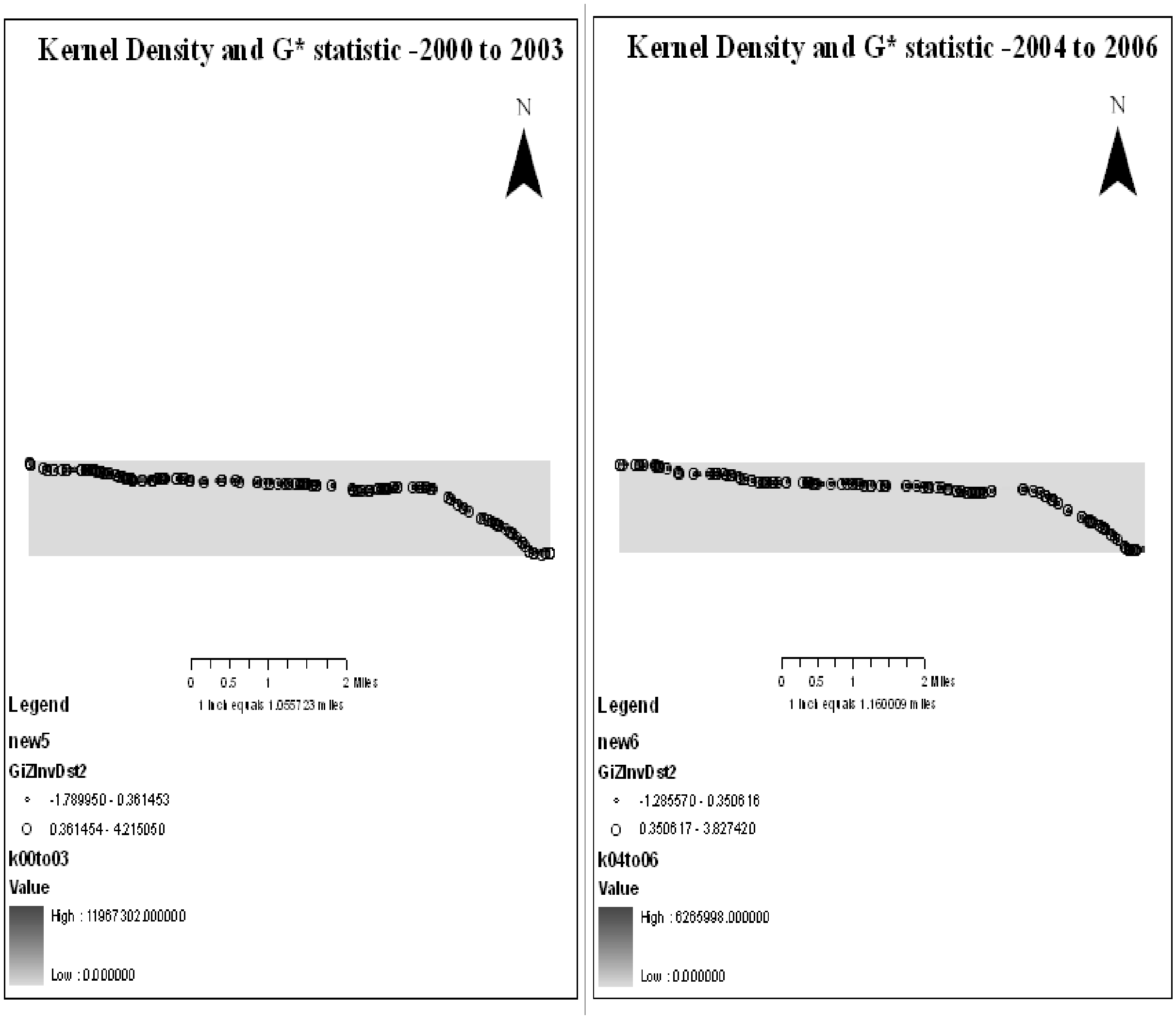
(c)

Figure 5. Hotspot identification by three CSR for G_i* statistic and K for year 2003

Figures 2 to 5 indicate hotspot identification using Getis-Ord G_i^* statistic and the Kernel density estimation (K) by each year from 2000 to 2003. From Figures 2 to 5 (a) represent the use of fixed distance, (b) represent the use of inverse distance, and (c) represent the use of inverse square distance, as the CSR for Getis-Ord G_i^* statistic for the identification of hotspots. In the figures presented above, design (boxes, triangles, etc.) represent the hotspots identified by the G_i^* statistic and dots shaded represent hotspots identified by K. It should be noted in the current study that only high values (among high and low) were compared between G_i^* statistic and K. The analysis was carried out by projecting the crashes in ArcGIS (ver. 9.2), and the tools for K and G_i^* were used. For the fixed distance band, a threshold value (proximity) of 100 feet was chosen. For inverse distance and inverse squared distance, the proximity was set to 'zero' to indicate that all features were considered neighbors to all other features. Additionally, the crashes were given weights of 542, 29, 11, 6 and 1 based on the severity level. The weights are based on the crash costs for each crash type (Highway Safety Manual, 2010). They are the ratios of different crash costs with respect to property damage only crashes. The costs are $4,008,900 for fatal crashes (S1); $216,000 for major injury (S2); $79,000 for minor injury (S3); $44,900 for complain of pain (S4) and $7,400 for property damage only crashes (S5), i.e., weight for fatal crash is 4008900/7400 = 542. Similarly, the other weights were computed and rounded to whole numbers.

Results from Figures 2 to 5 indicate that the identification of hotspots by K and G_i^* including three CSR for hotspots (high) were similar. Additionally, the three CSR methods identified similar hotspots (high). Further, the range of Z values for G_i^* for the hotspots (high) were same for the three CSR's used for years 2000 and 2001. For 2002 and 2003, range of Z values for G_i^* for the hotspots (high) were identical for inverse distance and inverse square distance CSR's. The identification of hotspots by inverse distance and inverse square distance CSR were similar and higher than the fixed distance CSR. These results show that the inverse distance and inverse distance square identified hotspots accurately and they were similar to K. Additional testing on the threshold distance was not carried out for the fixed distance CSR.

The crash data was aggregated from 2000 to 2003 and 2004 to 2006, and the hotspots were identified by inverse square distance CSR of G_i^* and K. This was carried out to check if the aggregation of crash data would affect the results of hotspot identification. The additional data i.e., 2004 to 2006 was used to check if the hotspots are identified similarly for G_i^* and K. Though the results presented and discussed in this paper are limited to hotspots with high level, the results should be similar for hotspots with low level. The authors did not perform and present this analysis as the identification of hotspots with higher concentrations is of more interest to the safety community.



(a) (b)

Figure 6. Hotspot identification by three CSR of G_i* statistic and K for 2000-2003 and 2004-2006

Figures 6(a) and (b) indicate that the identification of hotspots were the same for K and G_i*, when inverse square CSR was used for G_i*. Two possible reasons for similar identification may be i) the factors contributing to crashes and ii) existence of a functional relationship between these two functions. However, it can be noted that for the two methods, the crash contributing factors like alignment, profile, atmospheric conditions like rain, clear, driver behavior like DUI were not considered. For both of these methods, only the spatial location of the crash and the effect of the severity were considered. To understand the mathematical relationship between the two functions, consider the following:

$$\text{Density} = (\text{mass})/(\text{volume}) \quad (4)$$

$$\text{Density} = (\text{mass})/(\text{area*length}) \quad (5)$$

If this is summed over an entire section; considering sections are very small:

$$\Sigma \text{ Density } (D) = \Sigma \text{ (mass/area)} * \Sigma \text{ (1/length)} \tag{6}$$

$$D = \sum \frac{y_i}{d_{ij}^2} * \sum \frac{1}{length} \tag{7}$$

where D = density, y_i is the mass , d_{ij} is the distance, and

$$G_i^* = \frac{\sum w_{ij} y_i}{\sum y_i} \tag{8}$$

If inverse square distance is used then $w_{ij} = M/(d_{ij}^2)$. where M is a constant.

Therefore,

$$G_i^* = M \frac{\sum (y_i / d_{ij}^2)}{\sum y_i} \tag{9}$$

$$G_i^* = \frac{M * N}{\sum y_i} \tag{10}$$

where: $N = \sum (y_i / d_{ij}^2)$

Hence, from Equations (7) and (10)

$$G_i^* = \left(\frac{D * \sum \frac{1}{length}}{\sum y_i}\right) * M \tag{11}$$

From Equation (11), it can be stated that G_i* depends on density. In this case, Kernel density estimation is a special case of density function. Therefore, G_i* statistic can be used for point pattern analysis for kernel density. The distance chosen is inverse squared distance and this is similar to choosing inverse distance. This indicates a plausible explanation for similar identification of hotspots by the Getis-Ord G_i* and the Kernel density estimation.

Crash Contributing Factors

Though the crash contributing factors do not affect the identification of hotspots by G_i* and K, some of the key factors can be discussed. From Table 1, it can be inferred that on an average 12% of crashes are due to curved alignment, and 23% are due to grade roadway profile conditions. These may be related to improper sight distance, inadequate signage, etc. When types of collisions were considered, rear-end crashes contribute to 54% of the crashes, followed by single-vehicle crashes with 23% and sideswipe-same-direction with 14% of the crashes. The reason can be the higher AADT with an average of 94,921 vehs/yr. I-630 is located in the Arkansas state capital of Little Rock with a population of 183,133 and an area of 116.8 square miles (US Census, 2008). High occurrence of rear-end and sideswipe same direction crashes are common in major cities with high volume of traffic. The higher AADT indicates higher crash frequency but when traffic flow becomes very high, and speeds decrease a lot, crashes (especially fatal) may decrease. This was observed in the current study as there were only five fatal crashes through the analysis period out of which four cases included DUI, light conditions were dark, and single vehicle crashes, including three on the weekends. Though the detailed analyses of the crash contributing factors were not analyzed, elaborate information based on the frequency of crashes in specific cases were listed for additional information of interested readers.

CONCLUSIONS AND RECOMMENDATIONS

The main objective of this paper is compare two statistical techniques, kernel density estimation (K) and Getis-Ord G_i^* statistic using a GIS for hotspot identification. These methods identified similar hotspots. These results are similar to the findings from Flahaut et al. (2003). However, they used different autocorrelation index, local indicator of spatial autocorrelation (LISA) and different roadway and crash characteristics. A possible explanation for similar identification of hotspots is the mathematical forms of the two functions G_i^* and K. However, this is under specific parameters considered. These methods require strong statistical foundation and accurate positioning of each crash. In specific cases, the use of G_i^* is sufficient to identify hotspots.

Additionally, three conceptualization of spatial relationships (CSR) were used to determine G_i^* namely, fixed distance band, inverse distance and inverse square distance and the best CSR for hotspot identification is recommended. Results indicate that the inverse distance and inverse distance square identified hotspots accurately and similar to K than the fixed distance CSR. Additional testing on the threshold distance was not carried out for the fixed distance CSR.

Some of the contributing factors for the occurrence of these crashes include higher percentage of rear-end crashes, sideswipe same direction which may be due to higher AADT. Other factors include curved roadway and graded roadway profile which may be due to improper sight distance, inadequate signage, etc. Though analysis of the crash contributing factors is not the scope of the present study, based on the data presented only some of the factors were discussed.

This paper is limited to one highway, 7.4 miles long along a single county. The categorization of the hotspots was carried out in only two levels (high and low). Further analyses should include longer road sections and more than two levels. It should be noted that this paper is limited to exploratory data analysis of crashes. Further research aims in identifying the hotspots by developing models and including the effects of spatial autocorrelation.

ACKNOWLEDGMENTS

The authors are grateful to the Arkansas Highway and Transportation Department for providing data for this research and four anonymous reviewers for providing useful comments.

REFERENCES

Affum, J. K., and Taylor, A. P. (1995). Technology Tools for Transportation Professionals—Moving Into the 21st Century. *Resource Papers for the 1995 International Conference, Integrated GIS Database for Road Safety Management.* Institute of Transportation Engineers, Washington, D.C., pp. 189–193.

ArcGIS 9.2 Desktop Help Online Manual http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Standard_classification_schemes Accessed, June 23, 2010. ArcGIS 9.2 Desktop Help Online Manual.

Austin, K., Tight, M., and Kirby, H. (1997). The Use of Geographical Information Systems to Enhance Road Safety Analysis. Transportation Planning and Technology, Vol. 20, No. 4, pp. 249–266.

Bailey, T. C., and Gatrell, A. C. (1995). Interactive spatial data analysis. Essex: Longman.

Campbell, J., and Knapp, K. (2005). Alternative Crash Severity Ranking Measures and the Implications on Crash Severity Ranking Procedures. Mid-Continent Transportation Research Symposium. Ames, Iowa.

Depue, L. Safety Management Systems: A Synthesis of Highway Practice. NCHRP Synthesis 322, Transportation Research Board of the National Academies, Washington, D.C., 2003.

Flahaut, B., Mouchart, M., Martin, E, S., and Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones A comparative approach. Accident Analysis and Prevention, Vol. 35, pp. 991-1004.

Getis, A., and Ord, J. K. The analysis of Spatial Association by use of Distance Statistics. Geographic Analysis, Vol. 24, No. 3, 1992, pp. 189-206.

Hakkert, A. S., and Mahalel. D. (1978). Estimating the Number of Accidents at Intersections from a Knowledge of the Traffic flow on the Approaches. In Journal of Accident Analysis and Prevention, Vol. 10, pp. 69-79.

Hauer, E., Harwood, D. W., Council, F. M., and Griffith, M.S. (2002). Estimating safety by the Empirical Bayes Method: A Tutorial. In Transportation Research Record: Journal of the Transportation Research Board, No. 1784, TRB, Washington, D.C., pp. 126-131.

HSM, Highway Safety Manual, (2010). Section 4-30 and Table 4A-1. First Edition.

Krygier, J. B. http://go.owu.edu/~jbkrygie/krygier_html/geog_353/geog_353_lo/geog_353_lo07.html Accessed, June 23, 2010. J. B. Krygier: Geography lecture notes.

Kim, K., and Levine, N. (1996). Using GIS to Improve Highway Safety. Computers, Environment, and Urban Systems, Vol. 20, No. 4, pp. 289–302.

Levine, N., Kim, K., and Nitz, L. (1995). Spatial Analysis of Honolulu Motor Vehicle Crashes—Part I: Spatial Patterns. Accident Analysis and Prevention,Vol. 27, No. 5, pp. 663–674.

Levine, N., Kim, K., and Nitz, L. (1995). Spatial Analysis of Honolulu Motor Vehicle Crashes—Part II: Zonal Predictions. Accident Analysis and Prevention, Vol. 27, No. 5, pp. 675–685.

Lewis, D. J. (2010). http://danieljlewis.org/2010/06/07/jenks-natural-breaks-algorithm-in-python/ Accessed, June 25, 2010. Daniel J Lewis, University College London, Department of Geography.

McGuigan, D. R. D. (1981). The Use of Relationships between Road Accidents and Traffic Flow in “Black-Spot” Identification. Traffic Engineering and Control, pp. 448–453.

McMahon, P. A Quantitative and Qualitative Analysis of the Factors Contributing to Collisions between Pedestrians and Vehicles along Roadway Segments. Masters Project, Department of City and Regional Planning, University of North Carolina at Chapel Hill, NC. 1999.

Miller, J. S. What Value May Geographic Information Systems Add to the Art of Identifying Crash Countermeasures? Report VTRC 99–R13, 1999.

Mitra, S (2009). Spatial autocorrelation and Bayesian spatial statistical method for analyzing fatal and injury crash prone intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 2136, TRB, Washington, D.C., pp. 92-100.

NHTSA (2006). http://www-nrd.nhtsa.dot.gov/Pubs/810623.PDF, Accessed March. 10, 2010.

Norden, M., J. Orlansky and H. Jacobs. (1956). Application of Statistical Quality-Control Techniques to Analysis of Highway-Accident Data. Bulletin 117, HRB, National Research Council, Washington, D.C., 1956, pp. 17-31.

O’Sullivan, D., and Unwin, D. J. (2002). Geographic information analysis. Hoboken, New Jersey: John Wiley.

Ord, J. K., and Getis, A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographic Analysis, Vol. 27, No. 4, 1995, pp. 286-306.

Peled, A., Haj-Yehia, B., Hakkert, A.S. (1996). ArcInfo Based Geographical Information System for Road Safety Analysis and Improvement. http://www.esri.com/library/userconf/proc96/TO50/PAP005/P5.HTM, accessed November, 2009.

Powers, M. and Carson, J. (2004). Before-After Crash Analysis: A Primer for the Empirical Bayes Method Tutorial. A report prepared for the Research Section, Montana Department of Transportation.

Pulugurtha, S. S., Krishnakumar, V.K., and Nambisan, S.S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. In Journal of Accident Analysis and Prevention, Vol. 39, No. 4, pp. 800-811.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. London: Chapman Hall.

Songchitruksa, P., and Zeng, X. Getis-Ord Spatial Statistics for Identifying Hot Spots Using Incident Management Data. Proceedings of Transportation Research Board 89th Annual Meeting, Washington D. C., 2010.

Tobler, W. A computer movie simulating urban growth in the Detroit region. Economic Geography, Vol. 46, No. 9, 1970, pp. 234-240.

Xie, Z., and Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. Journal of Computer, Environment and Urban Systems, Vol. 32, pp. 396-406.