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 Gi* statistic using a Geographic Information System (GIS) for hotspot identification. The standardized Gi* 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 Gi* 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 Gi* 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 Gi* 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 Gi* (inverse square distance, CSR). Results indicated similar hotspots identified by both methods. The reason may be due to the mathematical functions of K and Gi*.

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; Songchitruxa 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* (Getis and Ord, 1992; Ord and Getis, 1995) statistics using a GIS for hotspot identification. The choice of choosing K and G* statistic is discussed in the later sections of the paper. Additionally, three conceptualization of spatial relationships (CSR) were used to determine G* 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*1), 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* 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*, is preferred for hotspot identification. Further specific details of G* 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.
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):

$$G_i^* = \sum_{j=1}^{n} w_{ij} x_j / \sum_{j=1}^{n} x_j$$

(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:

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

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:

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

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

* indicate the percentage of crashes
Figure 2. Hotspot identification by three CSR for $G_i^*$ statistic and $K$ for year 2000
Figure 3. Hotspot identification by three CSR for $G_i^*$ statistic and $K$ for year 2001
Figure 4. Hotspot identification by three CSR for $G_i^*$ statistic and K for year 2002
Figure 5. Hotspot identification by three CSR for $G^*_1$ 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 $4,008,900/7,400 = 542$. 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.
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:

\[
\text{Density} = \frac{\text{mass}}{\text{volume}}
\]

\[
\text{Density} = \frac{\text{mass}}{\text{area}\times\text{length}}
\]

If this is summed over an entire section; considering sections are very small:
\[ \Sigma \text{Density} (D) = \Sigma \left( \frac{\text{mass}}{\text{area}} \right) \times \Sigma \left( \frac{1}{\text{length}} \right) \] (6)

\[ D = \sum \frac{y_i}{d_{ij}^2} \times \sum \frac{1}{\text{length}} \] (7)

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

\[ G_i^* = \frac{\sum w_{ij} y_i}{\sum y_i} \] (8)

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

Therefore,

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

\[ G_i^* = \frac{M \times N}{\sum y_i} \] (10)

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

Hence, from Equations (7) and (10)

\[ G_i^* = \left( \frac{D \times \sum \frac{1}{\text{length}}}{\sum y_i} \right) \times M \] (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

Moving Into the 21st Century. Resource Papers for the 1995 International Conference,
Integrated GIS Database for Road Safety Management. Institute of Transportation

ArcGIS 9.2 Desktop Help Online Manual

Enhance Road Safety Analysis. Transportation Planning and Technology, Vol. 20, No. 4,
pp. 249–266.


Implications on Crash Severity Ranking Procedures. Mid-Continent Transportation


and the kernel method for identifying black zones A comparative approach. Accident


from a Knowledge of the Traffic flow on the Approaches. In Journal of Accident

Empirical Bayes Method: A Tutorial. In Transportation Research Record: Journal of the


Krygier, J. B.
http://go.owu.edu/~jbkrygie/krygier_html/geog_353/geog_353_lo/geog_353_lo07.html


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.


